2016-03

Relating tropical cyclone track forecast error distributions with measurements of forecast uncertainty

Chisler, Nicholas M.

Monterey, California: Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

RELATING TROPICAL CYCLONE TRACK FORECAST ERROR DISTRIBUTIONS WITH MEASUREMENTS OF FORECAST UNCERTAINTY

by

Nicholas M. Chisler

March 2016

Thesis Advisor: Wendell Nuss
Second Reader: Patrick Harr

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704–0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202–4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704–0188) Washington DC 20503.			
1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2016	**3. REPORT TYPE AND DATES COVERED** Master's thesis	
4. TITLE AND SUBTITLE RELATING TROPICAL CYCLONE TRACK FORECAST ERROR DISTRIBUTIONS WITH MEASUREMENTS OF FORECAST UNCERTAINTY		**5. FUNDING NUMBERS**	
6. AUTHOR(S) Nicholas M. Chisler			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943–5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A		**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number ____N/A____.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited		**12b. DISTRIBUTION CODE**	

13. ABSTRACT (maximum 200 words)

Tropical cyclone (TC) track forecasts will always contain uncertainty. This thesis relates ranges (bins) of uncertainty measurements with historical TC track forecast errors, to provide statistically distinct error distributions for use with the Monte Carlo (MC) method. T-test and Kolmogorov-Smirnov tests are used to confirm distinctness among error distributions associated with the bins of either European Center for Medium-Range Weather Forecasts (ECMWF) ensemble spread or TVCN Goerss Predicted Consensus Error (GPCE). The statistical tests indicate that distinct error distributions (consisting of official TC forecast error, ECMWF ensemble mean [EMN] error, or TVCN error) exist when using four bins of uncertainty (of either uncertainty measurement). Furthermore, error distributions of ECMWF EMN error are distinct with five bins of ECMWF ensemble spread. Along- and cross-track official errors could not be directly related to either measurement of uncertainty at even three bins. These results suggest that the National Hurricane Center test and evaluate the use of four bins of uncertainty for operational use with the MC method to further improve its Wind Speed Probability products and overall TC track forecasts. TC forecasters should also exploit the more impressive relationship established using five bins ECMWF ensemble spread with ECMWF EMN error.

14. SUBJECT TERMS tropical cyclone forecast, track error, uncertainty, Monte Carlo, ECMWF, TVCN, GPCE, ensemble spread, wind speed probability			**15. NUMBER OF PAGES** 75
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540–01–280–5500 Standard Form 298 (Rev. 2–89) Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

RELATING TROPICAL CYCLONE TRACK FORECAST ERROR DISTRIBUTIONS WITH MEASUREMENTS OF FORECAST UNCERTAINTY

Nicholas M. Chisler
Captain, United States Air Force
B.S., Purdue University, 2008

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN METEOROLOGY

from the

NAVAL POSTGRADUATE SCHOOL
March 2016

Approved by: Wendell Nuss
Thesis Advisor

Patrick Harr
Second Reader

Wendell Nuss
Chair, Department of Meteorology

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Tropical cyclone (TC) track forecasts will always contain uncertainty. This thesis relates ranges (bins) of uncertainty measurements with historical TC track forecast errors, to provide statistically distinct error distributions for use with the Monte Carlo (MC) method. T-test and Kolmogorov-Smirnov tests are used to confirm distinctness among error distributions associated with the bins of either European Center for Medium-Range Weather Forecasts (ECMWF) ensemble spread or TVCN Goerss Predicted Consensus Error (GPCE). The statistical tests indicate that distinct error distributions (consisting of official TC forecast error, ECMWF ensemble mean [EMN] error, or TVCN error) exist when using four bins of uncertainty (of either uncertainty measurement). Furthermore, error distributions of ECMWF EMN error are distinct with five bins of ECMWF ensemble spread. Along- and cross-track official errors could not be directly related to either measurement of uncertainty at even three bins. These results suggest that the National Hurricane Center test and evaluate the use of four bins of uncertainty for operational use with the MC method to further improve its Wind Speed Probability products and overall TC track forecasts. TC forecasters should also exploit the more impressive relationship established using five bins ECMWF ensemble spread with ECMWF EMN error.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I. INTRODUCTION 1

II. BACKGROUND 3
- **A. NATIONAL HURRICANE CENTER'S WIND SPEED PROBABILITY PRODUCT 3**
- **B. DESCRIPTION OF ENSEMBLES 5**
  - **1. European Centre for Medium-Range Weather Forecasts Ensemble 5**
  - **2. Consensus Models 7**
- **C. GOERSS PREDICTED CONSENSUS ERROR 8**
- **D. COMBINING THE MC METHOD WITH THE GPCE 9**
- **E. OTHER RELEVANT WORK 10**
- **F. GOALS OF THESIS 10**

III. METHODOLOGY 13
- **A. DATA 13**
  - **1. TIGGE 13**
  - **2. ATCF 14**
    - ***a. A-Decks 14***
    - ***b. B-Decks 14***
    - ***c. E-Decks 14***
- **B. DATA QUALITY CONTROL 15**
  - **1. Unrepresentative Errors 15**
  - **2. Ensuring Proper Data Pool 15**
    - ***a. Two-sample t-Test 15***
    - ***b. Two-sample KS-Test 15***
    - ***c. Testing Results 16***
- **C. BINNING BY ESTIMATED UNCERTAINTY 18**
  - **1. Constructing the Bins 18**
  - **2. Checking for Unique Error Distributions 18**
- **D. MAXIMIZING THE NUMBER OF BINS 19**
- **E. EXAMINING ALONG- AND CROSS-TRACK ERRORS 19**
- **F. COMPARING MODEL ERROR WITH MEASUREMENTS OF UNCERTAINTY 20**

IV. ANALYSIS AND RESULTS 21

A. RELATING UNCERTAINTY MEASUREMENTS WITH OFFICIAL FTE 21
1. ECMWF Ensemble Spread 21
2. TVCN GPCE Radii 21
B. ESTABLISHING BINS WITH MEASUREMENTS OF UNCERTAINTY 23
1. Using ECMWF Ensemble Spread to Establish Bins 24
2. Using TVCN GPCE Radii to Establish Bins 26
C. RESULTS OF STATISTICAL TESTING 29
1. ECMWF Ensemble Spread versus Official FTE 29
2. TVCN GPCE Radii versus Official FTE 34
D. ADDITIONAL FINDINGS 38
1. Using ATE and XTE 38
a. ECMWF Ensemble Spread vs. Official ATE 41
b. ECMWF Ensemble Spread vs. Official XTE 41
c. TVCN GPCE Radii vs. Official ATE 42
d. TVCN GPCE Radii vs. Official XTE 42
2. ECMWF Ensemble Spread vs. ECMWF EMN Error 42
3. TVCN GPCE Radii vs. TVCN Error 47

V. CONCLUSION AND RECOMMENDATIONS 51
A. CONCLUSIONS 51
B. RECOMMENDATIONS 52

APPENDIX. RANGES FOR BINS OF UNCERTAINTY 53

LIST OF REFERENCES 55

INITIAL DISTRIBUTION LIST 57

LIST OF FIGURES

Figure 1. WSP Product–Text Version 4

Figure 2. WSP Product–Graphical Version 5

Figure 3. Anomaly Correlations of 500 hPa Height 5-Day Forecasts from Several Operational Numerical Prediction Models 7

Figure 4. ECMWF Ensemble Spread vs. Official FTE at 60 Hours 22

Figure 5. TVCN GPCE Radius vs. Official FTE at 60 Hours 23

Figure 6. Three Bins of ECMWF Ensemble Spread 25

Figure 7. Four Bins of ECMWF Ensemble Spread 25

Figure 8. Five Bins of ECMWF Ensemble Spread 26

Figure 9. Three Bins of TVCN GPCE Radii 27

Figure 10. Four Bins of TVCN GPCE Radii 28

Figure 11. Five Bins of TVCN GPCE Radii 28

Figure 12. ECMWF Ensemble Spread vs. Official ATE at 60 Hours 39

Figure 13. ECMWF Ensemble Spread vs. Official XTE at 60 Hours 39

Figure 14. TVCN GPCE Radii vs. Official ATE at 60 Hours 40

Figure 15. TVCN GPCE Radii vs. Official XTE at 60 Hours 40

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1. Statistical Testing of ECMWF EMN Error: 2015 vs. Other Years 17

Table 2. Statistical Testing of ECMWF EMN Error: 2014 vs. Other Years 17

Table 3. Statistical Testing of TVCN Error: 2015 vs. Other Years 17

Table 4. Statistical Results of ECMWF Spread vs. Official FTE (3 Bins).....31

Table 5. Statistical Results of ECMWF Spread vs. Official FTE (4 bins) 32

Table 6. Statistical Results of ECMWF Spread vs. Official FTE (5 bins) 33

Table 7. Statistical Results of TVCN GPCE Radii vs. Official FTE (3 Bins).....35

Table 8. Statistical Results of TVCN GPCE Radii vs. Official FTE (4 Bins).....36

Table 9. Statistical Results of TVCN GPCE Radii vs. Official FTE (5 Bins).....37

Table 10. Statistical Results of ECMWF Ensemble Spread vs. EMN Error (3 Bins).....44

Table 11. Statistical Results of ECMWF Ensemble Spread vs. EMN Error (4 Bins).....45

Table 12. Statistical Results of ECMWF Ensemble Spread vs. EMN Error (5 Bins).....46

Table 13. Statistical Results of TVCN GPCE Radii vs. TVCN Error (3 Bins).....48

Table 14. Statistical Results of TVCN GPCE Radii vs. TVCN Error (4 Bins).....49

Table 15. Statistical Results of TVCN GPCE radii vs. TVCN Error (5 Bins) 50

Table 16. ECMWF Ensemble Spread Ranges (3 Bins) 53

Table 17. ECMWF Ensemble Spread Ranges (4 Bins) 53

Table 18. ECMWF Ensemble Spread Ranges (5 Bins) 53

Table 19. TVCN GPCE Radii Ranges (3 Bins) 54

Table 20. TVCN GPCE Radii Ranges (4 Bins) 54

Table 21. TVCN GPCE Radii Ranges (5 Bins) 54

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ATCF	Automated Tropical Cyclone Forecast
ATE	along-track error
AVNI	interpolated GFS
CONU	consensus model comprised of at least two of: GFDI, AVNI, NGPI, UKMI, and GFNI
ECMWF	European Center for Medium Range Weather Forecasting
EGRI	interpolated UKMET
EMXI	interpolated ECMWF
FTE	total-track error
GFDI	interpolated GFDL
GFDL	Geophysical Fluid Dynamics Laboratory
GFNI	interpolated Navy GFDL
GFS	Global Forecast System
GHMI	interpolated GFDL
GPCE	Goerss Predicted Consensus Error
hPa	hectopascal
HWFI	interpolated HWRF
HWRF	Hurricane Weather Research and Forecast System
KS	Kolmogorov-Smirnov
MC	Montel Carlo
NHC	National Hurricane Center
NGPI	adjusted NGPS
NGPS	Navy Operational Global Prediction System
NWP	numerical weather prediction
TC	tropical cyclone
THORPEX	The Observing System Research and Predictability Experiment
TIGGE	THORPEX Interactive Grand Global Ensemble
TS	tropical storm
TVCN	consensus model comprised of at least two of: GFDI, AVNI, NGPI, UKMI, and GFNI

UKMET	United Kingdom Meteorological
UKMI	interpolated UKMET
WSP	wind speed probability
XTE	cross-track error

ACKNOWLEDGMENTS

I would like to thank numerous people who were critical to the creation of this thesis. First off, Dr. Wendell Nuss provided me the overall guidance and direction to keep me on track during this process. His insight, Fortran coding skills, and willingness to explain key concepts to me (sometimes numerous times) was invaluable. Dr. Patrick Harr also played a critical role in the development of this thesis. His notes and code from previous work saved an incalculable number of hours, and his guidance on the statistical testing was greatly appreciated. Mary Jordan was another key contributor to this thesis. Without her adept ability to code in MATLAB, I could not have completed this thesis on time. She was a dependable Wingman who was always there in a pinch. Finally, I would like to thank my wife, Jodi, and my son, Chase. Without you two keeping me motivated and taking care of everything else when I was overwhelmed, I would not have finished this program.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Accuracy in forecasting the tracks of tropical cyclones (TC) has greatly improved over the past decades. Despite these improvements, the consumers of such forecasts (e.g., Department of Defense, local officials, businesses, general public) demand even more accurate information. This demand is not unwarranted, considering that the potential costs of inaccurate TC track forecasts include the livelihood of millions of coastal dwelling citizens and an unfathomable value in lost resources, infrastructure, personal property, and lives. While the need is straightforward, the complexities and challenges to formulating accurate TC track forecasting are much less so.

There exist numerous sources of forecasting error that can be minimized, but cannot be eliminated: this is why weather forecasts are not (and never will be) perfect. These sources of error are embedded in the very tools—the observations, model physics, mathematical methods, and assumptions—that forecasters must use to make a forecast. We are, however, able to advance our ability to forecast effectively by progressively minimizing these sources of error, while also increasing our computational power.

Once we accept that error will always be present in a forecast, the next best solution besides eliminating the error is to quantify and characterize the error. Ensembles are a great tool for identifying where and when the inherent error will grow and manifest. Through utilizing many ensemble members based on varying initial conditions, parameterizations, etc., a range of possible outcomes is revealed to the forecaster. We assume that the truth lies somewhere within the range of outcomes, and that when the range is relatively small (large), there is a high (low) degree of certainty in the forecast.

This thesis aims to quantify the degree of uncertainty represented by an ensemble and relate it to TC track forecast error. If successful, this relationship will allow forecasters to apply a unique range of possible forecast errors to each individual TC. The benefit of such a relationship will be more representative TC track forecasts and associated wind speed probability (WSP) products from the National Hurricane Center (NHC). Improvements to these products will not only provide the consumers with the

best forecast possible, but also relay the level of uncertainty unique to a given storm. This will better inform decision makers to help protect all assets at risk.

II. BACKGROUND

A. NATIONAL HURRICANE CENTER'S WIND SPEED PROBABILITY PRODUCT

The National Hurricane Center (NHC) has produced probability products since the early 1980s; however, a significant advancement in such products was implemented in 2006 (DeMaria et al. 2009). The NHC's TC WSP product incorporates uncertainties in track, intensity, and wind structure. According to DeMaria et al., a Monte Carlo (MC) method is utilized to give the probability of winds reaching or exceeding 34, 50, and 64 kt at a given time and location. A random sample of 1,000 errors is drawn using the MC method from a distribution of official track and intensity errors based on the most recent five years of data. These samples are then added to the official forecast to produce 1,000 realizations (plausible forecasts). Probabilities that wind speeds will reach a given threshold can then be calculated by identifying how many of the realizations reach the threshold for a given time and location (2009).

Figures 1 and 2 are examples of the WSP product in text and graphical form, respectively. The NHC explains that the text product provides two types of probabilities, onset and cumulative, for each location listed. The former refers to the probability that the threshold will be met during the specific time window, while the latter refers to the probability that the threshold will be met at any time up to that forecast hour. The graphical form of the product only informs the user of the cumulative probability that the given threshold will be met at any time up to the given forecast hour (NHC 2014). While the graphical form only gives one probability type and does not include exact percentages, it enables the user to see the approximate probabilities anywhere on the map.

These products will benefit from the research reported in this thesis through the refinement of the distributions from which the 1,000 realizations produced by the MC method are drawn. By drawing from a refined set of errors, the realizations will adjust with the nature of the uncertainty for any given storm. Thus, the probabilities may

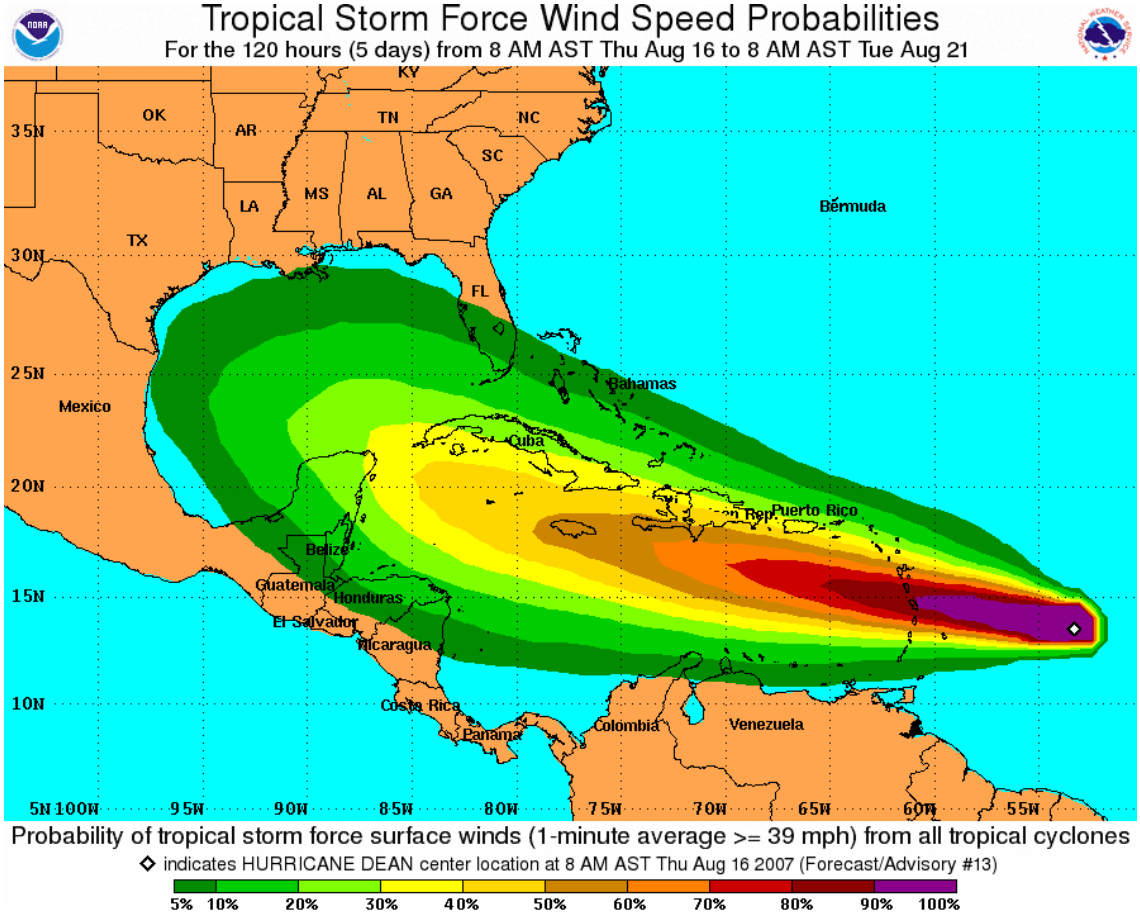
increase or decrease (and the probability swath may widen or narrow) to provide a more customized TC track forecast.

Figure 1. WSP Product–Text Version

```
TROPICAL STORM ARTHUR WIND SPEED PROBABILITIES NUMBER   7
NWS NATIONAL HURRICANE CENTER MIAMI FL       AL012014
1500 UTC WED JUL 02 2014

AT 1500Z THE CENTER OF TROPICAL STORM ARTHUR WAS LOCATED NEAR
LATITUDE 29.1 NORTH...LONGITUDE 79.1 WEST WITH MAXIMUM SUSTAINED
WINDS NEAR 50 KTS...60 MPH...95 KM/H.

Z INDICATES COORDINATED UNIVERSAL TIME (GREENWICH)
   ATLANTIC STANDARD TIME (AST)...SUBTRACT 4 HOURS FROM Z TIME
   EASTERN  DAYLIGHT TIME (EDT)...SUBTRACT 4 HOURS FROM Z TIME
   CENTRAL  DAYLIGHT TIME (CDT)...SUBTRACT 5 HOURS FROM Z TIME

WIND SPEED PROBABILITY TABLE FOR SPECIFIC LOCATIONS

CHANCES OF SUSTAINED (1-MINUTE AVERAGE) WIND SPEEDS OF AT LEAST
   ...34 KT (39 MPH... 63 KM/H)...
   ...50 KT (58 MPH... 93 KM/H)...
   ...64 KT (74 MPH...119 KM/H)...
FOR LOCATIONS AND TIME PERIODS DURING THE NEXT 5 DAYS

PROBABILITIES FOR LOCATIONS ARE GIVEN AS OP(CP) WHERE
    OP  IS THE PROBABILITY OF THE EVENT BEGINNING DURING
        AN INDIVIDUAL TIME PERIOD (ONSET PROBABILITY)
   (CP) IS THE PROBABILITY OF THE EVENT OCCURRING BETWEEN
        12Z WED AND THE FORECAST HOUR (CUMULATIVE PROBABILITY)

PROBABILITIES ARE GIVEN IN PERCENT
X INDICATES PROBABILITIES LESS THAN 1 PERCENT
PROBABILITIES FOR 34 KT AND 50 KT ARE SHOWN AT A GIVEN LOCATION WHEN
THE 5-DAY CUMULATIVE PROBABILITY IS AT LEAST 3 PERCENT.
PROBABILITIES FOR 64 KT ARE SHOWN WHEN THE 5-DAY CUMULATIVE
PROBABILITY IS AT LEAST 1 PERCENT.


  - - - - WIND SPEED PROBABILITIES FOR SELECTED LOCATIONS - - - -

               FROM    FROM    FROM    FROM    FROM    FROM    FROM
  TIME       12Z WED 00Z THU 12Z THU 00Z FRI 12Z FRI 12Z SAT 12Z SUN
PERIODS         TO      TO      TO      TO      TO      TO      TO
             00Z THU 12Z THU 00Z FRI 12Z FRI 12Z SAT 12Z SUN 12Z MON

FORECAST HOUR    (12)   (24)    (36)    (48)    (72)    (96)   (120)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
LOCATION       KT

HIBERNIA OILFD 34  X   X( X)   X( X)   X( X)   X( X)   X( X)   4( 4)

CAPE RACE NFLD 34  X   X( X)   X( X)   X( X)   X( X)   3( 3)   8(11)

ILE ST PIERRE  34  X   X( X)   X( X)   X( X)   X( X)  11(11)   8(19)
ILE ST PIERRE  50  X   X( X)   X( X)   X( X)   X( X)   1( 1)   3( 4)

BURGEO NFLD    34  X   X( X)   X( X)   X( X)   X( X)  15(15)  13(28)
BURGEO NFLD    50  X   X( X)   X( X)   X( X)   X( X)   2( 2)   3( 5)

PTX BASQUES    34  X   X( X)   X( X)   X( X)   X( X)  23(23)  11(34)
PTX BASQUES    50  X   X( X)   X( X)   X( X)   X( X)   4( 4)   4( 8)
PTX BASQUES    64  X   X( X)   X( X)   X( X)   X( X)   1( 1)   X( 1)
```

This truncated version of the text form of the WSP product provides two probabilities: onset and cumulative. The onset probability (the first of each pair of numbers) gives the likelihood that the threshold will be met during that specific time window. The cumulative probability (indicated with parenthesis) gives the likelihood that the threshold will be met at any time up to that forecast hour. Adapted from NHC, 2014: Tropical cyclone wind speed probabilities products. Accessed on 19 January 2016. [Available online at http://www.nhc.noaa.gov/pws_example.shtml.]

Figure 2. WSP Product–Graphical Version



The graphical version of the WSP product gives the cumulative probability that the given threshold will be met at any time up to the given forecast hour. Source: NHC, 2014: Tropical cyclone wind speed probabilities products. Accessed on 19 January 2016. [Available online at http;//www.nhc.noaa.gov/gifs/WindSpeedProbGraphic.gif.]

B. DESCRIPTION OF ENSEMBLES

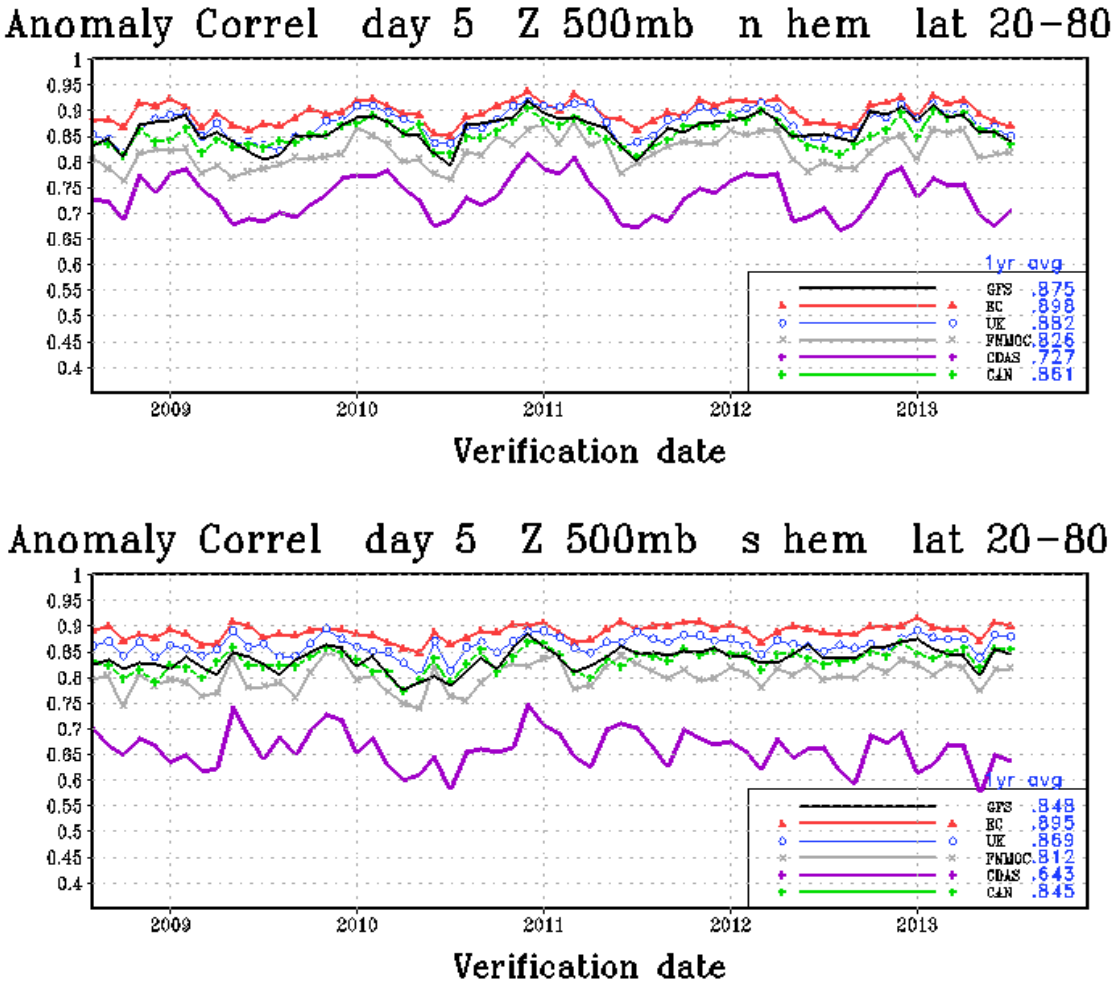
The primary purpose of ensemble forecast systems is to quantify uncertainty in a forecast. This can be accomplished through single-model ensembles or multi-model ensembles (consensus models), both of which NHC has access to.

1. European Centre for Medium-Range Weather Forecasts Ensemble

The European Centre for Medium-Range Weather Forecasts (ECMWF) global-model ensemble is comprised of 51 members. Fifty of the members are created using a

slightly perturbed variation of the ECMWF analysis while the fifty-first member is created using the original analysis and a coarser resolution than the deterministic forecast. This ensemble is a powerful tool for estimating forecast uncertainty via examination of the spread and/or grouping of all the members' forecasts. The global ensemble forecast system is the basis for many products related to midlatitude and tropical circulation systems. For this thesis, ensemble spread is defined as the average distance from the forecast TC position in each member to the ensemble mean (EMN) forecast TC position. The ECMWF is commonly recognized as the most accurate weather model available. Figure 3 is a comparison of six of the top weather models: GFS (Global Forecast System), ECMWF (EC in the legend of Figure 3), UKMET (United Kingdom Meteorology), FNMOC (Fleet Numerical Meteorology and Oceanography Center), CDAS (Climate Data Assimilation System), and CMC (Canadian Meteorological Centre). The data shows that the ECMWF consistently outperforms the other models at predicting day 5 500 hPa heights from 2009–2013.

Figure 3. Anomaly Correlations of 500 hPa Height 5-Day Forecasts from Several Operational Numerical Prediction Models



A comparison of six leading weather models: GFS, ECMWF, UKMET, FNMOC, CDAS, and CMC. The data above shows how well each model has performed at forecasting day 5 500 hPa anomalous heights from 2009–2013 for both the northern Hemisphere (top) and southern Hemisphere (bottom). A y-value of 1 represents a perfect forecast as compared to analyzed heights. Source: NCEP: Accessed 22 February 2016. [Available online at http://www.emc.ncep.noaa.gov/gmb/STATS/html/aczhist6.html.]

2. Consensus Models

Consensus models are utilized based on the idea that the average of two or more imperfect models will, on average, be more accurate than any single model forecast. This is similar to the concept of creating an ensemble from a single model by running it many times with slightly perturbed initial conditions. However, a notable difference between

these two approaches to making an ensemble is that consensus models may be comprised with as few as two members (ranging up to approximately 5–7), while single-model ensembles frequently have dozens of members. The TVCN[1] is a consensus model frequently used by the NHC to help predict the track of TCs and is used in this thesis along with the ECMWF ensemble.

C. GOERSS PREDICTED CONSENSUS ERROR

Another tool to utilize consensus models was created by Goerss (2007) to help identify and quantify forecast track uncertainty. The Goerss predicted consensus error (GPCE) provides a way to statistically estimate consensus model error (DeMaria 2013). According to Goerss, the tool works by taking into account numerous parameters such as ensemble spread, initial and forecast TC intensity, initial TC position, and forecast displacement. Of these parameters, Goerss found ensemble spread to be the most important, followed by initial and forecast TC intensity (Goerss 2007).

Goerss established relationships between the aforementioned parameters and consensus TC track error. He utilized the consensus model defined as CONU[2] in his 2007 study. Goerss then established a procedure by which forecast CONU TC track errors could be utilized with stepwise linear regression-based parameters to establish forecasts for each forecast hour (2007).

Finally, using the predicted CONU TC forecast errors derived from the linear regression models, combined with varying constants for each forecast hour, Goerss created GPCE circles. The circles are defined by a radius based on spread of model forecasts in CONU and centered at the forecast position for each forecast hour of the CONU. These circles were designed so that they would contain the verifying TC position ~70% of the time (Goerss 2007). Therefore, this circle provides an estimate of forecast

[1] TVCN is comprised of five models: GHMI (interpolated GFDL [Geophysical Fluid Dynamics Laboratory]), EGRI (interpolated UKMET with subjective quality control), HWFI (interpolated HWRF [Hurricane Weather Research and Forecast System]), AVNI (interpolated GFS), and EMXI (interpolated ECMWF model) E. Hendricks, personal communication, March 1, 2016).

[2] CONU is a consensus model comprised of at least two of the following models: GFDI (interpolated GFDL), AVNI, NGPI (adjusted NGPS [Navy Operational Global Prediction System]), UKMI (interpolated UKMET), and GFNI (interpolated Navy GFDL).

uncertainty similar to that obtained from the ensemble spread of a single-model ensemble. Since Goerss's work in 2007, the GPCE has been implemented at the NHC using the TVCN consensus model and is how this study incorporates the GPCE as an error estimate.

D. COMBINING THE MC METHOD WITH THE GPCE

As previously described, the MC method is used to draw 1,000 samples from the previous 5-year official forecast error distribution of track and intensity. While this technique was beneficial, it treated all TCs and their forecast errors as equal. In other words, the forecast track errors of all prior TCs in the past 5 years were grouped together to create a distribution, randomly drawn from, and then applied to each new TC that formed in the corresponding basin. However, by utilizing the GPCE, TCs (past and present) can be grouped together to form bins with distinct characteristics based on their estimated forecast uncertainty.

Hauke (2006) researched the possibility of binning TC errors into terciles based on the TC forecast GPCE value (calculated using the CONU consensus model) for a given forecast hour. The resulting terciles represented TC forecasts with low, average, and high degrees of uncertainty. Hauke's work also examined the possibility of using the GFS ensemble spread as the parameter to create the three bins (Hauke, 2006). The goal of his work was to discover whether the error distributions associated with each of the three bins (for both approaches) were significantly different. If so, then that would establish a more unique error distribution for the MC method to draw from to provide a more refined wind probability distribution. His studies utilized errors calculated from the official track forecasts produced by NHC and investigated this potential relationship using the total track errors (FTE), along-track errors (ATE), and cross-track errors (XTE) (Hauke, 2006). ATE is defined as the component of the FTE that is parallel to the storm track. Positive (negative) errors represent forecast positions ahead (behind) of actual TC position. XTE is defined as the component of the FTE that is perpendicular to the storm track. Positive (negative) errors represent forecast positions to the right (left) of the actual TC position.

According to Hauke, binning TCs into terciles using GPCE values proved to create statistically different error distributions in all three categories (FTE, ATE, and XTE). These results suggest that the MC method would benefit from such stratification. However, the results of binning TCs into terciles using GFS ensemble spread was not as successful. Hauke (2006) concluded that such stratification would not benefit the MC method and may even degrade its performance.

The lack of skill in using GFS ensemble estimates of uncertainty may be due in part to limitations in the size of the GFS ensemble (21 members) or in how it is perturbed. However, due to the impressive track record of the ECMWF and the increased ensemble size relative to the GFS, this thesis aims to utilize its estimates of uncertainty to accomplish what could not be established using the GFS ensemble.

E. OTHER RELEVANT WORK

A crucial element to improving the MC method through the use of uncertainty information is to provide statistically distinct error distributions from which to draw. The uncertainty-skill relationship is thought to vary due to differing storm characteristics. While Hauke (2006) stratified by magnitude of uncertainty, Neese (2010) attempted to stratify by storm location (sub regions) within the Atlantic basin. While Neese's results were inconclusive, his work suggests the possibility that benefit may be attained by binning error distributions based on TC location.

Next, Pearman (2011) studied the effectiveness of using a GPCE ellipse that contained both along- and cross-track uncertainty estimates instead of the GPCE circle. Pearman used a Grand Ensemble (combination of multiple ensembles) to provide the data for his work. While the results of his work indicate that the GPCE ellipse performs as well (if not better in some instances) as the GPCE circle, this method still remains experimental today (Pearman 2011).

F. GOALS OF THESIS

The goal of this thesis is to extend the uncertainty-skill relationship by examining a longer data set encompassing multiple years of TC forecasts. Presumably, the bin size

of uncertainty can be further refined to establish a greater number of bins producing statistically different error distributions as measured by distribution mean and shape. A larger number of discrete bins allows for a more continuous relationship between forecast error and estimated uncertainty to be derived. This process was repeated for two different measures of uncertainty: ECMWF ensemble spread and the GPCE radius as calculated from the TVCN. This study used the official forecast error to represent skill in three different ways: FTE, ATE, and XTE. Finally, each measure of uncertainty was used in conjunction with its corresponding model's forecast error to establish their ability to predict the parent model's error as opposed to official track error, which utilizes objective guidance. If these results are significantly different compared those found using official errors, then applying the wind probability model to that forecast might be more helpful.

This thesis provides a technique to be used in conjunction with MC method that will provide unique error distributions for multiple levels of uncertainty as conveyed by the ECMWF ensemble or TVCN GPCE radius. This process allows for a more tailored forecast for each new TC, and should result in improvements to NHC's WSP products.

THIS PAGE INTENTIONALLY LEFT BLANK

III. METHODOLOGY

The overall approach used in this study was to calculate TC forecast track errors for the official NHC forecasts as well as two models and examine the errors relative to estimates of forecast uncertainty. Specifically, the ECMWF ensemble, TVCN consensus model, and NHC official forecasts were used. The ECMWF ensemble spread and TVCN GPCE radius provide the uncertainty estimates.

A. DATA

The data analyzed in this thesis spans the years 2007 through 2015, and come from all forecasts for TCs that occurred over the Atlantic basin during that time. That includes 123 named storms (67 tropical storms [TS] and 56 hurricanes). Omitted TCs include Hurricane Noel and TS Olga (’07), TS Marco (’08), and TS Nicole (’10) due to missing or incomplete data. Data were retrieved from The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) and the NHC Automated Tropical Cyclone Forecast (ATCF) system. These data systems provided the historical official track forecasts and accompanying ECMWF and TVCN model data for this study.

1. TIGGE

All of the ECMWF ensemble data were obtained via TIGGE, an archive of ensemble forecast data from ten global numerical weather prediction (NWP) centers that is used primarily for scientific research (Santoalla, 2015). Specifically, the forecast from each of the available 50 members of the ECMWF ensemble is included. Each member of the ensemble provides a TC forecast (if applicable) for each forecast run during the TC’s existence. The ECMWF ensemble is run at 0000 UTC and 1200 UTC. The TIGGE data include the storm name, TC position (latitude and longitude), central pressure, and wind speed at each forecast hour (12, 24, 36, 48, 60, 72, 84, 96, 108, and 120).

2. ATCF

The ATCF data set is produced operationally by NHC. The data are contained in three files named: A-Decks, B-Decks, and E-Decks. These decks contain both forecast verification and guidance products for each TC.

a. A-Decks

The A-Decks contain the official TC track and intensity forecasts along with other NWP guidance. The official forecasts are those which are created and distributed by the NHC. The official forecasts are provided every six hours and include the forecast hours of 3, 12, 24, 36, 48, 72, 96, and 120 hours. The official forecasts do not contain the 60, 84, and 108 forecast hours, but they are derived through interpolation to allow for comparisons with other products. These official forecasts provide the basis from which the WSP products are created. For the purpose of this work, they are used with the B-Decks to establish the official FTE.

b. B-Decks

The B-Decks contain the best track (verified) positions for each TC. The best track position is determined during the post-storm analysis. It takes into account all relevant information that may not have been available during the storm for inclusion in analyses and forecasts. The best track data includes storm number, position, central pressure, and wind speed every six hours. All track errors in this study were calculated from the forecast position compared to the best track positions.

c. E-Decks

The E-Decks contain guidance used to provide a measure of confidence in the track forecast consensus aids. For the purpose of this thesis, the GPCE associated with the TVCN will be utilized. The GPCE value is the radius of a circle that is calculated to contain the true TC position ~70% of the time. TVCN is run every six hours and provides information at all applicable forecast hours (12, 24, 36, 48, 60, 72, 96, and 120). The E-Decks are also interpolated to obtain the 84 and 108 forecast hours.

B. DATA QUALITY CONTROL

1. Unrepresentative Errors

After data retrieval, a filter was applied to exclude any data that originated from a time when the TC was not categorized at a tropical storm (winds ≥ 34 kt) or hurricane (winds ≥ 64 kt). The purpose of this filter is to reduce cases where a TC center may be subjective or indistinct. Such TCs may lead to forecasts with unrepresentative errors that would pollute the data sample. In addition, TCs that became extratropical were not included once the transition occurred.

2. Ensuring Proper Data Pool

While a large data pool is desired for statistical work, the data samples need to be examined to ensure that they are not statistically different. Given that models evolve over time, the forecast skill from one year to a later year could be substantially different. In other words, we had to ensure that data from each year (2007–2015) were similar enough to be pooled for analysis. In order to accomplish this, two statistical tests where utilized: a two-sample t-test and a two-sample Kolmogorov-Smirnov test (KS-test). Each test was performed with the data from 2015 compared with the data from each of the prior years.

a. Two-sample t-Test

A two-sample t-test is used to determine whether the means of two samples are statistically different from each other. It assumes that each sample follows a Gaussian (normal) distribution. In this thesis, the null hypothesis of the t-test is that the means of the data samples (years) are not significantly different. Thus by testing all of the years against 2015, we can see if our entire sample is statistically the same or if advancements in the models over the years have caused the data to become statistically different. The version of the t-test utilized in this study uses a 95% confidence level and assumes unequal variances of the samples.

b. Two-sample KS-Test

A two-sample KS-test is used to determine whether the distributions of two samples are statistically the same. This test evaluates the uniqueness of the shape of each

distribution rather than just the means of the distribution as in the t-test. In this thesis, the null hypothesis of the KS-test is that the distributions of both data samples are statistically the same. This is the same as saying that both data samples are drawn from the same distribution. A 95% confidence level was also utilized with the KS-test.

c. Testing Results

After all of the tests were conducted, it became apparent that 2015 contained unusually low errors, as calculated using the ECMWF EMN forecast. This resulted in rejected null hypotheses for many comparisons, particularly for 2010 and 2011. This suggests that there are differences in the model performance from year to year when compared to 2015. In order to examine whether this variability was caused by a characteristic of the 2015 sample, the same testing was conducted using the data from 2014 compared with the data from each of the prior years. Results from all of the testing for both the ECMWF and TVCN models are shown in Tables 1, 2, and 3. Green cells indicate that the null hypothesis failed to be rejected, while red cells indicate the null hypothesis was rejected.

While it appears that 2010 and 2011 have statistically different ECMWF EMN errors at first, the use of a second set of comparisons in Table 2 shows that all of the years are more similar than not. Keeping in mind that this relatively small number of years may have outliers that skew the errors of certain years, this data set appears to be similar enough throughout the years and forecast hours to be grouped together and certainly did not exhibit any systematic trend in performance over time. Consequently, the full set of years was used for all subsequent analysis in this study.

Table 1. Statistical Testing of ECMWF EMN Error: 2015 vs. Other Years

T-Test and KS-Test Results for 2015 ECMWF Error vs. Other Years																
Fcst	2014		2013		2012		2011		2010		2009		2008		2007	
Hour	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test
12	0	0	1	0	1	1	1	0	1	1	0	0	0	0	1	1
24	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	1
36	0	0	0	0	0	0	1	0	1	1	0	0	0	0	1	1
48	0	0	0	1	0	0	1	1	1	1	0	1	1	0	1	1
60	0	0	0	0	1	0	1	1	1	1	0	0	1	0	0	0
72	0	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0
84	0	0	0	0	1	0	1	1	1	1	0	0	1	0	0	0
96	1	0	0	0	1	0	1	1	1	1	0	0	1	0	0	0
108	1	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0
120	1	1	0	0	1	0	1	1	1	1	0	0	0	0	0	0

Results of statistical testing of ECMWF EMN error comparing 2015 to all other years. Green cells indicate a failure to reject the null hypothesis. Red cells indicate a rejection of the null hypothesis.

Table 2. Statistical Testing of ECMWF EMN Error: 2014 vs. Other Years

T-Test and KS-Test Results for 2014 ECMWF Error vs. Other Years														
Fcst	2013		2012		2011		2010		2009		2008		2007	
Hour	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test
12	1	0	0	0	1	0	1	0	1	0	1	0	1	1
24	0	0	0	0	0	0	1	1	0	1	0	0	1	1
36	0	0	0	0	1	0	1	1	0	0	0	0	1	1
48	0	0	0	0	0	0	1	0	0	0	0	0	0	0
60	0	0	0	0	1	0	1	1	0	0	0	0	0	0
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0
84	0	0	0	0	1	0	1	0	0	0	0	0	0	0
96	0	0	0	0	0	0	0	0	0	0	0	0	0	0
108	0	0	0	0	0	0	0	0	0	0	0	1	0	0
120	0	1	0	1	0	0	0	0	1	1	1	1	0	0

Results of statistical testing of ECMWF EMN error comparing 2014 to all other years. Green cells indicate a failure to reject the null hypothesis. Red cells indicate a rejection of the null hypothesis.

Table 3. Statistical Testing of TVCN Error: 2015 vs. Other Years

T-Test and KS-Test Results for 2015 TVCN Error vs. Other Years														
Fcst	2014		2013		2012		2011		2010		2009		2008	
Hour	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test	T-Test	KS-Test
12	0	0	0	0	0	0	0	0	0	0	0	1	0	0
24	0	0	0	0	0	0	1	0	1	0	0	0	0	0
36	0	0	0	1	0	0	0	0	1	0	0	0	1	0
48	0	0	0	0	0	0	0	0	1	1	0	0	0	0
60	0	0	0	0	0	0	0	0	1	0	0	0	0	0
72	0	0	0	0	0	0	0	0	1	0	0	0	0	0
84	0	0	0	0	0	0	0	0	0	0	0	0	0	0
96	1	0	0	0	0	0	0	0	0	0	0	0	0	0
108	0	1	0	0	0	0	0	0	0	0	0	0	0	0
120	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Results of statistical testing of TVCN EMN error comparing 2015 to all other years. Green cells indicate a failure to reject the null hypothesis. Red cells indicate a rejection of the null hypothesis.

C. BINNING BY ESTIMATED UNCERTAINTY

A goal of this study is to examine the forecast uncertainty to forecast error relationship. While a continuous relationship is desired, the approach used to get there is to subdivide the uncertainty into discrete bins and test for the statistical uniqueness of their associated error distributions. If those subdivisions prove to provide unique error distributions, then increase the number of subdivisions until statistical uniqueness is lost.

1. Constructing the Bins

While Hauke (2006) previously showed that binning by terciles (three bins) of forecast uncertainty proved beneficial for CONU GPCE, this study begins with three bins as well for the ECMWF ensemble spread and TVCN GPCE in order to confirm these previous results based on data from one year. In order to establish the three bins, all of the pairs of uncertainty measures and corresponding official forecast errors were arranged from least to greatest uncertainty for each forecast hour. The values of uncertainty that correspond to one-third and two-thirds of the data population were used as the cutoff values to create the three bins. The goal was to create three bins with an equal number of data points; however, that would have required splitting up a set of data points with the same measurement of uncertainty into different bins for some cases. To avoid this, the bins are close to being equal but are not exactly equal for all forecast hours. The exact ranges for all bins established in this work are in Tables 16–21 and can be found in the Appendix.

2. Checking for Unique Error Distributions

After the bins were established, the next step was to check the error distributions for uniqueness. To accomplish this, the t-test and KS-test were utilized again. The only change from the variations of the statistical tests utilized for the year to year comparison is related to the t-test. In order to compare the distributions of each bin, a right-tailed version of the t-test was used. This version of the test only checks if the mean of the second sample (e.g., Bin 2) is greater than that of the first sample (e.g., Bin 1). The right-tailed t-test was chosen because we are assuming that each progressive bin will have a larger mean than the previous. Both tests were then used to compare the first tercile

(lowest uncertainty) to the second. The tests were performed again between the second and third (highest uncertainty) bins. For this use of the statistical tests, the desired result was to reject the null hypotheses. That would indicate that the different bins of uncertainty have statistically different means and/or distributions, and thus can be used independently by NHC in the MC method.

D. MAXIMIZING THE NUMBER OF BINS

Having verified that the three bins of uncertainty estimates produce unique error distributions, the next step was to repeat this process using progressively more bins. Bins are identified using a range of 1 to N (where N = total number of bins). Bin 1 always represents the least amount of uncertainty, while bin N represents the greatest level of uncertainty as given by either ensemble spread or GPCE value. N level of bins were created using the same principles as described for three bins.

Finally, the bins were all compared using the t-test and KS-test again. The format for comparing the bins was as follows: bin 1 vs. bin 2, bin 2 vs. bin 3, … , bin N-1 vs. bin N. This process of increasing the number of bins was repeated until the sample of errors within the bins lacked statistical difference in their means and distribution. At this point, the data set was not sufficiently robust to draw meaningful conclusions regarding finer ranges of uncertainty.

E. EXAMINING ALONG- AND CROSS-TRACK ERRORS

Another possible relationship that yields benefit is comparing the distributions of official forecast ATE and XTE versus each of the measures of uncertainty. This process was nearly the same as that of the official FTE approach. The only differences are that the error distribution for each bin represented either ATE or XTE, and the two-tailed t-test was utilized. The reason for changing to the two-tailed t-test is because the assumption that the mean of each successive bin will increase is no longer valid. In fact, we expect the mean to stay near zero given the fact that there should be an approximately equal number of positive errors as negative errors.

The goal of performing these two additional sets of comparisons was to further fine tune the information that can be extracted from historical error data. Relationships defined between ATE and XTE versus an uncertainty measurement provides forecasters even more detailed information regarding the uncertainty. Specifically, it helps separate the uncertainty in the track of the TC from uncertainty in the speed of the TC.

F. COMPARING MODEL ERROR WITH MEASUREMENTS OF UNCERTAINTY

Finally, two more relationships worth analyzing are those between the ECMWF EMN error and spread, and between the TVCN error and GPCE radius. While relating measurements of uncertainty with official FTE provides the most directly relevant information to NHC, making the same relationships with model error provides additional information about model performance that is useful to forecasters. By establishing these relationships, forecasters can opt to modify either of these model outputs by applying the MC method to the model output using its own unique error distributions. These relationships were investigated using the techniques as described above. First, three bins were established and then the number of bins were maximized to extract the finest ranges of uncertainty that the data sample would allow to contain statistically unique means and distributions.

IV. ANALYSIS AND RESULTS

A. RELATING UNCERTAINTY MEASUREMENTS WITH OFFICIAL FTE

1. ECMWF Ensemble Spread

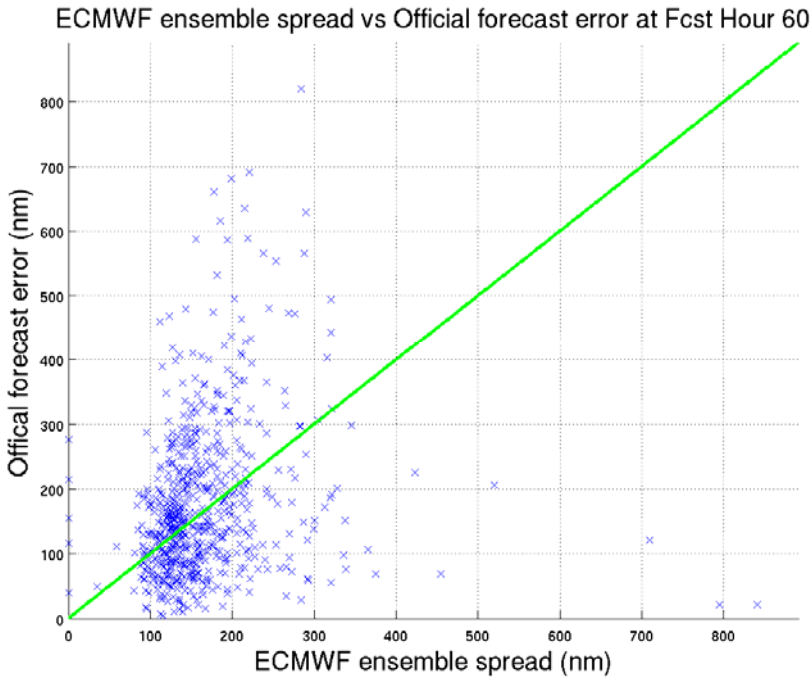
In general, we expect that when uncertainty in a forecast increases, the error associated with that forecast will increase, too (Scherrer, 2002). To demonstrate this relationship, the ECMWF ensemble spread is plotted versus official FTE at forecast hour 60 in Figure 4. Note that forecast hour 60 was chosen to demonstrate the expected patterns for all forecast hours in the following figures. This forecast hour is in the middle of the total forecast period and provides a good representation of error characteristics in the other forecast hours unless otherwise noted. The diagonal green line represents a one-to-one direct relationship between spread and FTE. While such an exact relationship is clearly not present, a trend of increasing FTE with increasing spread can be identified with the exception of a handful of outliers representing very high (low) spread with very low (high) FTE. This relationship does not illustrate a one-to-one relationship; instead the relationship is much steeper (i.e., very little variation in spread corresponds to large variations in FTE). The FTE tends to increase quickly with small increases in spread. While Figure 4 is valid for forecast hour 60, the other forecast hours display similar relationships. The range of spread and errors tend to be lower (higher) at the shorter (longer) forecast hours. The presence of such a trend indicates that a relationship likely exists between the uncertainty measurement and official FTE. Although not examined in this study, the extreme outliers may represent instances with a greatly reduced number of contributing ensemble members.

2. TVCN GPCE Radii

The relationship between TVCN GPCE radii and official FTE is shown in Figure 5. This relationship is very similar to that of ECMWF ensemble spread and official FTE. Once again, the relationship is one in which very little variation in spread corresponds to large variations in FTE. A notable difference between Figure 4 and Figure 5 is the range of the uncertainty measurement. The range of the GPCE values is significantly smaller

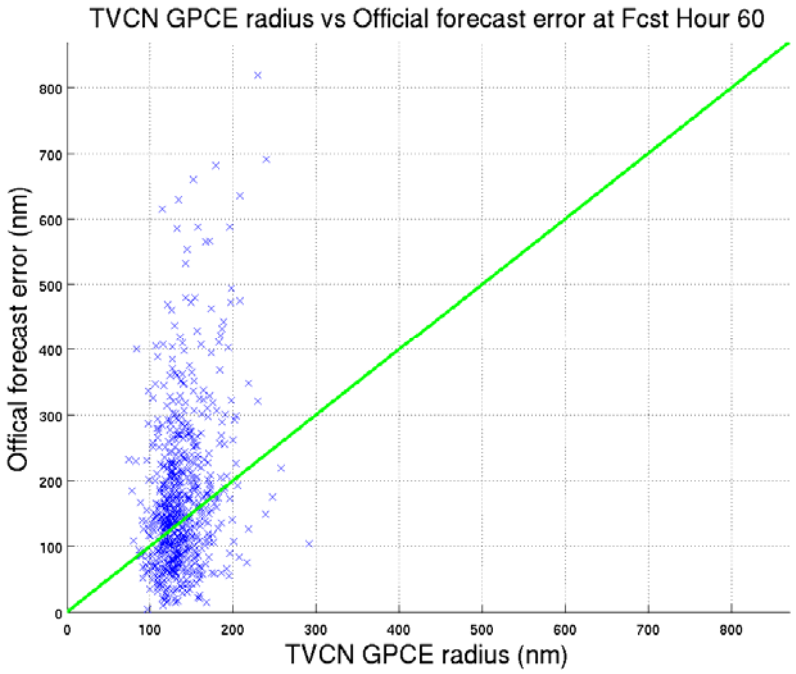
than that of the spread, and contains considerably fewer outliers. This lack of extreme outliers in the GPCE certainly reflects a less-sensitive measure of uncertainty as the variability in the multi-model consensus may not be as extreme as the ECMWF ensemble, or perhaps the GPCE calculation itself limits variability.

Figure 4. ECMWF Ensemble Spread vs. Official FTE at 60 Hours



A scatterplot of ECMWF ensemble spread vs. official FTE with one-to-one line (solid green).

Figure 5. TVCN GPCE Radius vs. Official FTE at 60 Hours



A scatterplot of TVCN GPCE radii vs. official FTE with a one-to-one line (solid green).

B. ESTABLISHING BINS WITH MEASUREMENTS OF UNCERTAINTY

In order to exploit the relationship found between the measurements of uncertainty and official FTE, the data were divided into N bins as noted in the methodology section. Each bin was plotted as a histogram to show the distribution of official FTE for each range of uncertainty. The error distributions must be significantly different from the next (as evaluated by the t-test and KS-test) for maximum benefit to be gained from breaking the data into N ranges of uncertainty.

When multiple bins of uncertainty are created, it is expected that the errors associated with each progressive bin will follow the general trend of increasing error size and variability. Specifically, as bin number increases (increasing forecast uncertainty), one expects the mean and standard deviation of the error distributions to increase. Another expected trend is that the number of small official FTE decreases with increasing uncertainty while the tail of the distribution (representing larger FTE) grows.

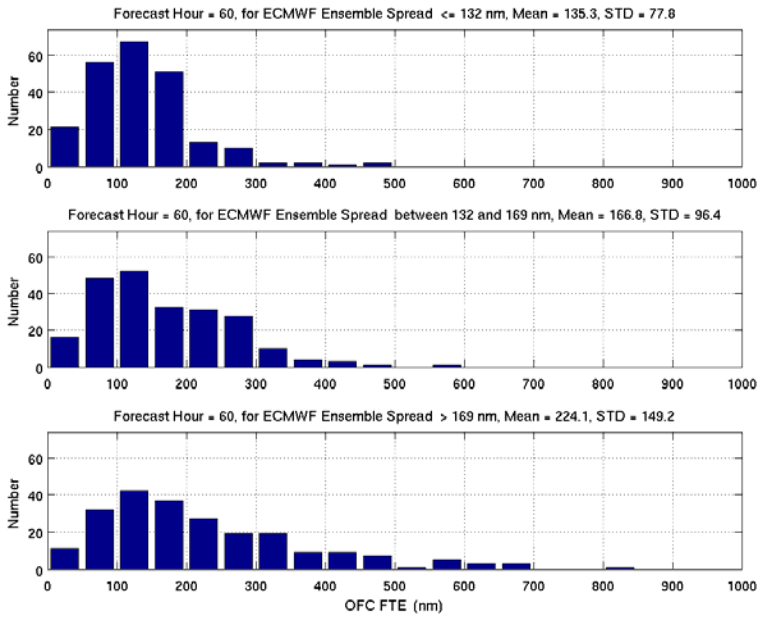
1. Using ECMWF Ensemble Spread to Establish Bins

Histograms that correspond to the three bins of ECMWF ensemble spread are shown in Figure 6. The top histogram represents forecasts that contained the lowest measurements of uncertainty (≤ 132 nm in this case) while the middle and lower histograms show forecasts with successively larger measurements of uncertainty. The three histograms appear to represent error distributions similar to what is expected. As the level of uncertainty increases, the mean and standard deviation of the error distributions increase as well. The mean increases from approximately 135 to 167 to 224 nm while the standard deviation increases from 78 to 96 to 149 nm. Through visual inspection, these plots suggest that they each represent a unique error distribution with different means and variance. A similar pattern can be seen in each of the other forecast hours which are not shown.

Histograms corresponding to the four bins of ECMWF ensemble spread are shown in Figure 7. With four ranges of uncertainty, the data still behave as expected in general. The mean increases from 129 to157 to 187 to 229 nm while the standard deviation increases from 76 to 85 to 112 to 157 nm. However, the distinctness becomes slightly less apparent through visual inspection, especially when comparing bins one and two.

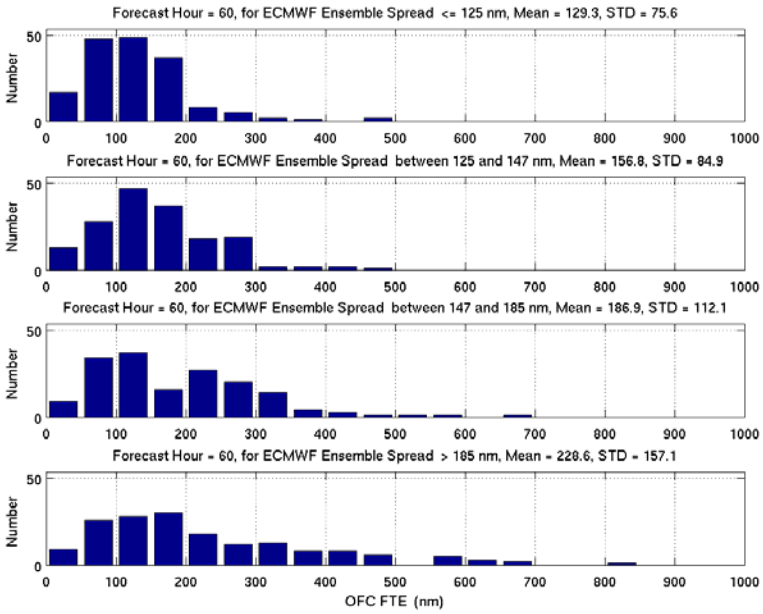
Histograms that correspond to the five bins of ECMWF ensemble spread are shown in Figure 8. With five bins, visual inspection begins to reveal similarities between multiple bin comparisons. The distributions between bins one and two appear quite similar as well as that between bins four and five. However, the mean still increases with each bin of uncertainty from 123 to 151 to 175 to 190 to 238 nm while the standard deviation increases from 72 to 80 to 102 to 120 to 160 nm. Although the mean and standard deviation still change as expected, a point will come where their changes in value will not be sufficient to be deemed statistically unique distributions. This result happens because the sample size of each bin becomes too small to draw meaningful conclusions. In this case, the sample sizes have been reduced to 135 data points for each bin.

Figure 6. Three Bins of ECMWF Ensemble Spread



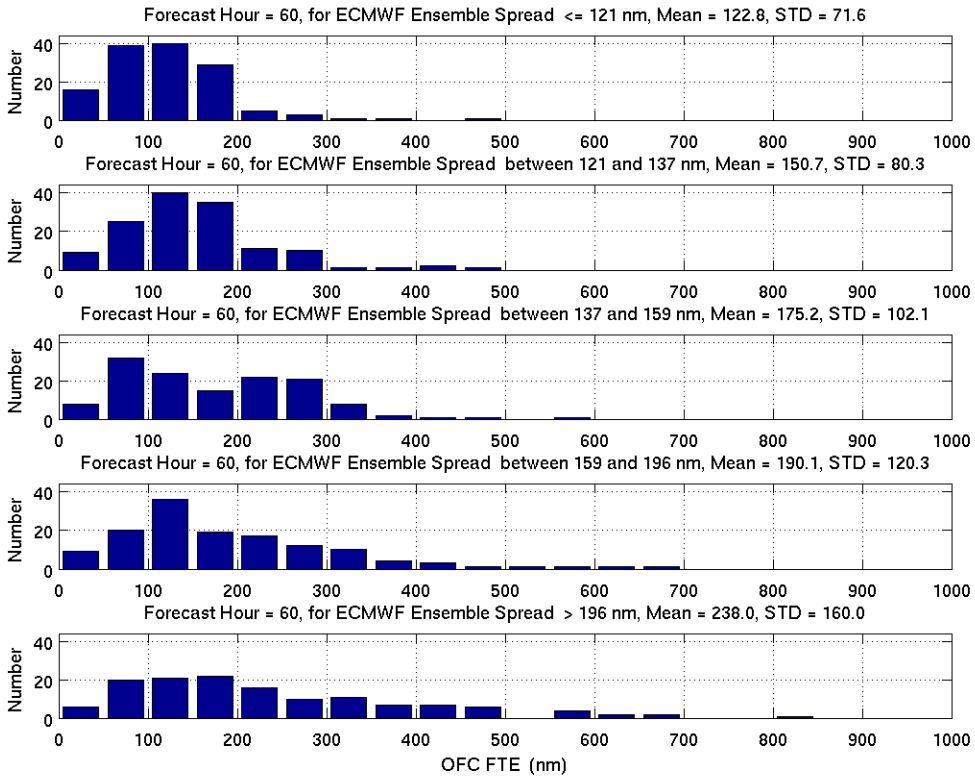
Histograms created by dividing ECMWF into three bins show official FTE distributions for each range of uncertainty.

Figure 7. Four Bins of ECMWF Ensemble Spread



Histograms created by dividing ECMWF into four bins show official FTE distributions for each range of uncertainty.

Figure 8. Five Bins of ECMWF Ensemble Spread



Histograms created by dividing the ECMWF ensemble spread into five bins show official FTE distributions for each range of uncertainty.

2. Using TVCN GPCE Radii to Establish Bins

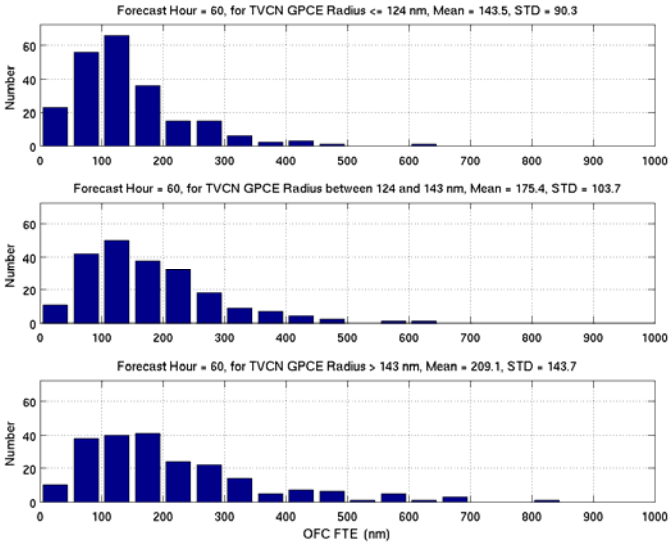
Histograms corresponding to the three bins of TVCN GPCE radii are shown in Figure 9. These histograms share the same characteristics as those created using ECMWF ensemble spread. The mean of each successive bin increases from 144 to 175 to 209 nm while the standard deviation increases from 90 to 104 to 144. Each of the histograms appears to belong to a distinct error distribution where the mean and standard deviation increases with higher levels of uncertainty. This same pattern can be seen in each of the other forecast hours not shown.

Histograms that correspond to the four bins of TVCN GPCE radii are shown in Figure 10. These error distributions also begin to become less visually distinct. The comparison of bins two and three begins to reveal similarities in the location and shape of

the error distributions. However, the means still increase from 140 to 163 to 179 to 222 nm while the standard deviation increases from 91 to 98 to 108 to 154 nm. Thus the distributions still display enough distinctness to proceed with five bins. While the means and standard deviations of the bins still increase with uncertainty, the increases become smaller and two of the bins begin to resemble each other, especially at the midrange forecast hours.

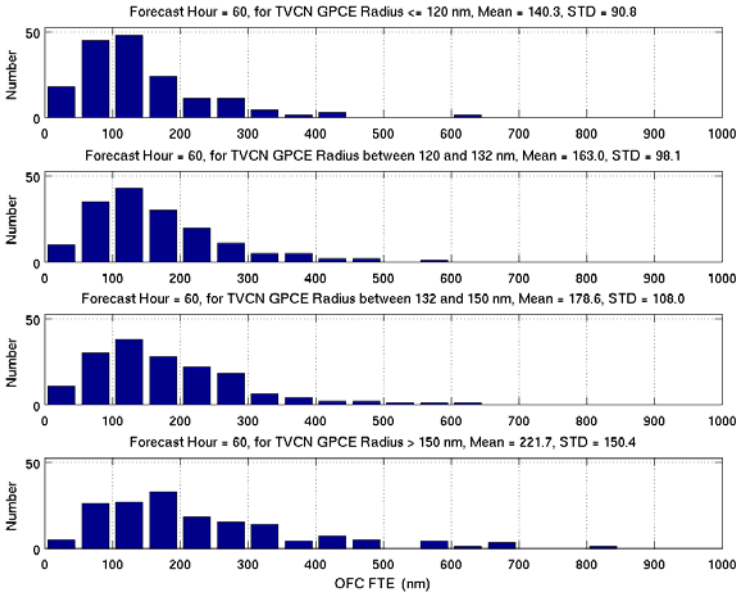
Histograms corresponding to the five bins of TVCN GPCE radii are shown in Figure 11. While these distributions maintain slightly more useful distinction than those created using five bins of ECMWF ensemble spread (Figure 8), the trend is still to become less distinctive. This pattern of becoming less distinct is observed in many of the other forecast hours as well. It is common for two sets of bins to begin taking on similar values for the mean and standard deviation. The mean of each bin increases from 144 to 154 to 168 to 187 to 227 nm while the standard deviation changes from 98 to 88 to 103 to 119 to 154 nm. Note that the outlier in bin one caused the standard deviation to be larger than that of bin two.

Figure 9. Three Bins of TVCN GPCE Radii



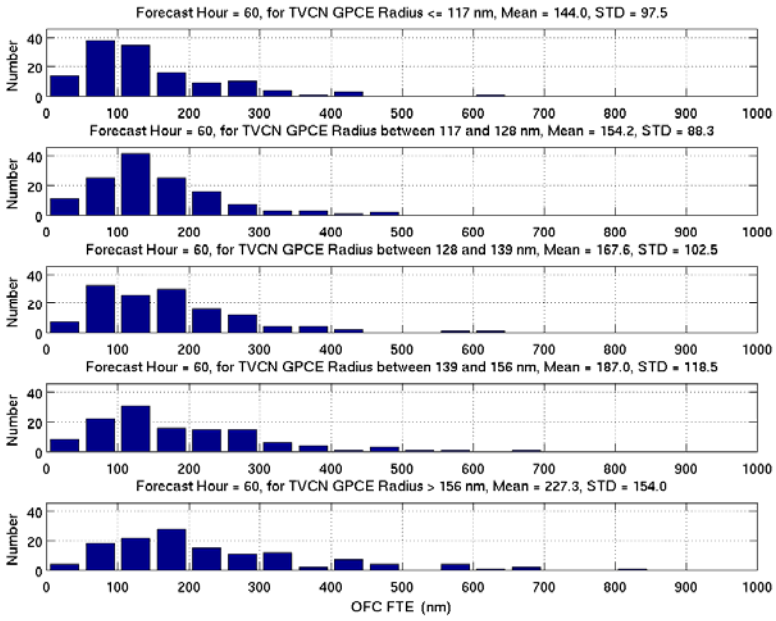
Histograms created by dividing the TVCN GPCE radii into three bins show official FTE distributions for each range of uncertainty.

Figure 10. Four Bins of TVCN GPCE Radii



Histograms created by dividing the TVCN GPCE radii into four bins show official FTE distributions for each range of uncertainty.

Figure 11. Five Bins of TVCN GPCE Radii



Histograms created by dividing the TVCN GPCE radii into five bins show official FTE distributions for each range of uncertainty.

C. RESULTS OF STATISTICAL TESTING

After all of the bins were established and the associated histograms where created and compared in the previous section, the data from each bin were statistically compared with each neighboring bin to check for distinctness. While visual inspection is a good first approximation of determining distinctness between two distributions, the t-test and KS-test use all the data points to determine if the null hypothesis should be rejected (H = 1). If not, the tests will result in a failure to reject the null hypothesis (H = 0). The null hypothesis for the t-test is that the mean of the two samples drawn from assumed Gaussian distributions are equal. The t-test's alternative hypothesis is that the mean of the second sample drawn from the assumed Gaussian distribution is greater than that of the first sample. The null hypothesis for the KS-test is that the errors of both bins came from the same distribution, while the alternative hypothesis is that the errors of both bins come from different distributions.

1. ECMWF Ensemble Spread versus Official FTE

Results from the statistical testing when dividing ECMWF ensemble spread into three bins are shown in Table 4. A quick glance at the table reveals that nearly every comparison at every forecast hour rejects each test's null hypothesis. This indicates that the associated Gaussian error distributions of each bin are deemed to have different means (t-test), and each sample (bin) of data comes from different distributions (KS-test) at a 95% confidence level. These results confirm the idea that TC forecasts can be subdivided into three ranges of ECMWF ensemble spread, each with its own distinct error distribution that the MC method can draw from to better relay tailored uncertainty information for any given TC.

Results from the statistical testing when dividing ECMWF ensemble spread into four bins are shown in Table 5. The results from these tests are not quite as concrete considering that 14 of the 60 tests failed to reject the null hypothesis. However, only three pairs of tests did so for the same bin comparison and forecast hour, thus there is still significant benefit to be attained by establishing four bins. When the t-test's null hypothesis is rejected, and the KS-test's null hypothesis fails to be rejected (8 of the 14

failed tests), valuable information is still obtained by comparing the individual distributions. This situation indicates that while the shape (variance) of the two distributions may not be different, their means are. Two error distributions of equal shape but different means will still provide the MC method with different errors to draw from.

Results from the statistical testing when dividing ECMWF ensemble spread into five bins are shown in Table 6. The statistical analysis of dividing the ECMWF ensemble spread into five bins shows that 36 of the 80 tests fail to reject the null hypothesis. More importantly, 13 pairs of tests failed to reject the null hypothesis for the same bin comparison and forecast hour. This indicates that there was no benefit attained by adding the fifth bin at the corresponding forecast hours. For example, Table 6 it can be seen that the comparison of Bin 1 and Bin 2 includes both tests rejecting the null hypothesis for 6 of the 10 forecast hours. This implies that the MC method will not improve by drawing from different distributions at those forecast hours, because the associated error distributions are statically the same.

Table 4. Statistical Results of ECMWF Spread vs. Official FTE (3 Bins)

Fcst Hour	ECMWF Spread vs. Official Error (3 bins)			
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
12	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0
	T-stat: 3.374	KS-Stat: 0.15	T-stat: 2.318	KS-Stat: 0.01
	P: 0	P: 0.002	P: 0.01	P: 0.115
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
24	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.501	KS-Stat: 0.208	T-stat: 4.671	KS-Stat: 0.2
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
36	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 4.904	KS-Stat: 0.217	T-stat: 3.725	KS-Stat: 0.174
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
48	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 4.632	KS-Stat: 0.183	T-stat: 3.944	KS-Stat: 0.195
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
60	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.812	KS-Stat: 0.209	T-stat: 4.839	KS-Stat: 0.182
	P: 0	P: 0	P: 0	P: 0.001
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
72	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.499	KS-Stat: 0.199	T-stat: 5.736	KS-Stat: 0.224
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
84	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.893	KS-Stat: 0.211	T-stat: 5.376	KS-Stat: 0.239
	P: 0.002	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
96	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.099	KS-Stat: 0.213	T-stat: 4.934	KS-Stat: 0.219
	P: 0.001	P: 0	P: 0	P: 0.001
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
108	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.484	KS-Stat: 0.173	T-stat: 5.104	KS-Stat: 0.254
	P: 0.007	P: 0.019	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
120	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.435	KS-Stat: 0.19	T-stat: 4.877	KS-Stat: 0.254
	P: 0.008	P: 0.012	P: 0	P: 0

T-test and KS-test results comparing official FTE error distributions obtained via three bins of ECMWF spread. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

Table 5. Statistical Results of ECMWF Spread vs. Official FTE (4 bins)

Fcst Hour	ECMWF Spread vs. Official Error (4 bins)					
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
12	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0
	T-stat: 2.092	KS-Stat: 0.133	T-stat: 2.422	KS-Stat: 0.159	T-stat: 1.245	KS-Stat: 0.088
	P: 0.019	P: 0.029	P: 0.008	P: 0.005	P: 0.107	P: 0.311
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
24	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.311	KS-Stat: 0.17	T-stat: 2.754	KS-Stat: 0.159	T-stat: 3.146	KS-Stat: 0.192
	P: 0.011	P: 0.003	P: 0.003	P: 0.006	P: 0.001	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
36	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 0
	T-stat: 4.222	KS-Stat: 0.228	T-stat: 2.126	KS-Stat: 0.125	T-stat: 2.142	KS-Stat: 0.102
	P: 0	P: 0	P: 0.017	P: 0.076	P: 0.016	P: 0.22
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
48	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 0
	T-stat: 3.446	KS-Stat: 0.191	T-stat: 2.553	KS-Stat: 0.13	T-stat: 1.869	KS-Stat: 0.13
	P: 0	P: 0.002	P: 0.006	P: 0.075	P: 0.031	P: 0.075
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
60	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0
	T-stat: 3.152	KS-Stat: 0.189	T-stat: 2.771	KS-Stat: 0.174	T-stat: 2.806	KS-Stat: 0.141
	P: 0.001	P: 0.004	P: 0.003	P: 0.01	P: 0.003	P: 0.063
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
72	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 3.845	KS-Stat: 0.232	T-stat: 1.935	KS-Stat: 0.105	T-stat: 3.947	KS-Stat: 0.185
	P: 0	P: 0	P: 0.027	P: 0.334	P: 0	P: 0.008
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
84	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 2.121	KS-Stat: 0.204	T-stat: 1.916	KS-Stat: 0.14	T-stat: 3.713	KS-Stat: 0.202
	P: 0.017	P: 0.005	P: 0.028	P: 0.123	P: 0	P: 0.006
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
96	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 1.707	KS-Stat: 0.15	T-stat: 3.001	KS-Stat: 0.216	T-stat: 3.48	KS-Stat: 0.178
	P: 0.045	P: 0.106	P: 0.001	P: 0.005	P: 0	P: 0.032
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
108	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 0.797	KS-Stat: 0.132	T-stat: 2.618	KS-Stat: 0.178	T-stat: 3.599	KS-Stat: 0.268
	P: 0.213	P: 0.258	P: 0.005	P: 0.048	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
120	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: -0.46	KS-Stat: 0.077	T-stat: 3.603	KS-Stat: 0.245	T-stat: 2.984	KS-Stat: 0.196
	P: 0.677	P: 0.907	P: 0	P: 0.001	P: 0.002	P: 0.032

T-test and KS-test results comparing official FTE error distributions obtained via four bins of ECMWF spread. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

Table 6. Statistical Results of ECMWF Spread vs. Official FTE (5 bins)

Fcst Hour	ECMWF Spread vs. Official Error (5 bins)							
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
12	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 0	KS-Test: 0
	T-stat: 0.462	KS-Stat: 0.082	T-stat: 2.093	KS-Stat: 0.015	T-stat: 1.81	KS-Stat: 0.128	T-stat: 0.343	KS-Stat: 0.056
	P: 0.322	P: 0.541	P: 0.019	P: 0.037	P: 0.036	P: 0.089	P: 0.366	P: 0.93
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
24	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 1.425	KS-Stat: 0.118	T-stat: 1.95	KS-Stat: 0.152	T-stat: 2.407	KS-Stat: 0.174	T-stat: 2.024	KS-Stat: 0.152
	P: 0.078	P: 0.157	P: 0.026	P: 0.03	P: 0.008	P: 0.008	P: 0.022	P: 0.03
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
36	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0
	T-stat: 2.807	KS-Stat: 0.193	T-stat: 0.197	KS-Stat: 0.061	T-stat: 4.092	KS-Stat: 0.23	T-stat: 0.64	KS-Stat: 0.084
	P: 0.003	P: 0.004	P: 0.422	P: 0.914	P: 0	P: 0	P: 0.261	P: 0.6
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
48	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0
	T-stat: 3.542	KS-Stat: 0.207	T-stat: 1.692	KS-Stat: 0.116	T-stat: 2.617	KS-Stat: 0.195	T-stat: -0.03	KS-Stat: 0.077
	P: 0	P: 0.002	P: 0.046	P: 0.247	P: 0.005	P: 0.005	P: 0.51	P: 0.746
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
60	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 3.009	KS-Stat: 0.237	T-stat: 2.194	KS-Stat: 0.237	T-stat: 1.097	KS-Stat: 0.126	T-stat: 2.775	KS-Stat: 0.163
	P: 0.001	P: 0	P: 0.015	P: 0	P: 0.137	P: 0.219	P: 0.003	P: 0.049
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
72	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 2.42	KS-Stat: 0.172	T-stat: 1.954	KS-Stat: 0.121	T-stat: 1.926	KS-Stat: 0.136	T-stat: 2.879	KS-Stat: 0.176
	P: 0.008	P: 0.043	P: 0.026	P: 0.295	P: 0.028	P: 0.183	P: 0.002	P: 0.037
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
84	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.029	KS-Stat: 0.15	T-stat: 0.998	KS-Stat: 0.205	T-stat: 2.671	KS-Stat: 0.167	T-stat: 2.709	KS-Stat: 0.188
	P: 0.152	P: 0.157	P: 0.16	P: 0.017	P: 0.004	P: 0.084	P: 0.004	P: 0.036
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
96	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 0
	T-stat: 1.373	KS-Stat: 0.136	T-stat: 1.23	KS-Stat: 0.14	T-stat: 2.445	KS-Stat: 0.183	T-stat: 2.667	KS-Stat: 0.159
	P: 0.086	P: 0.287	P: 0.11	P: 0.252	P: 0.008	P: 0.058	P: 0.004	P: 0.134
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
108	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 0.005	KS-Stat: 0.121	T-stat: 1.135	KS-Stat: 0.132	T-stat: 2.802	KS-Stat: 0.209	T-stat: 2.893	KS-Stat: 0.23
	P: 0.498	P: 0.493	P: 0.129	P: 0.383	P: 0.003	P: 0.032	P: 0.002	P: 0.013
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
120	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0
	T-stat: -1.02	KS-Stat: 0.096	T-stat: 1.209	KS-Stat: 0.133	T-stat: 2.896	KS-Stat: 0.223	T-stat: 2.3	KS-Stat: 0.156
	P: 0.845	P: 0.816	P: 0.114	P: 0.433	P: 0.002	P: 0.027	P: 0.011	P: 0.243

T-test and KS-test results comparing official FTE error distributions obtained via five bins of ECMWF spread. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. Yellow cells indicate a P-value within 1% of threshold. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

2. TVCN GPCE Radii versus Official FTE

Results from the statistical testing when dividing TVCN GPCE radii into three bins are shown in Table 7. Similar to Table 4, the test results for both the t-test and KS-test only include two instances where the null hypothesis failed to be rejected across all bin comparisons and forecast hours (one of which was within 1% and highlighted in yellow). These results also confirm the idea that TC forecasts can be subdivided into three ranges of GPCE radii, each with its own distinct error distribution that the MC method can draw from to better relay tailored uncertainty information for any given TC.

Results from the statistical testing when dividing TVCN GPCE radii into four bins are shown in Table 8. These results are not quite as strong as those in Table 5 for ECMWF ensemble spread with 14 of the 60 tests failing to reject the null hypothesis (including two within 1%). While that part is the same as Table 5, the difference is that there were six pairs of tests that failed to reject the null hypothesis for the same bin comparison and forecast hour—four of which lie within the Bin 2 versus Bin 3 comparison. While less robust than the results from ECMWF ensemble spread, creating four bins of TVCN GPCE radii still provides enough distinctness across forecast hours and each bin comparison to be beneficial. It appears that this approach performs best for short forecast hours (≤ 36) and struggles some in the midrange forecast hours (48–72).

Results from the statistical testing when dividing TVCN GPCE radii into five bins are shown are shown in Table 9. As with ECMWF ensemble spread, the test results are discouraging with 40 of 80 tests failing to reject the null hypothesis. This includes 15 pairs of tests for the same bin comparison and forecast hour. With 50% of the tests failing to support any potential benefit of adding a fifth bin, it seems that four bins is again the maximum subdivision that produces distinct error distributions for this data set.

Table 7. Statistical Results of TVCN GPCE Radii vs. Official FTE (3 Bins)

Fcst Hour	GPCE Radius vs. Official Error (3 bins)			
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
12	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.691	KS-Stat: 0.142	T-stat: 5.467	KS-Stat: 0.232
	P: 0	P: 0.005	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
24	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 6.077	KS-Stat: 0.242	T-stat: 4.086	KS-Stat: 0.188
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
36	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 4.444	KS-Stat: 0.225	T-stat: 5.01	KS-Stat: 0.12
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
48	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.236	KS-Stat: 0.154	T-stat: 4.232	KS-Stat: 0.168
	P: 0.001	P: 0.005	P: 0	P: 0.002
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
60	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0
	T-stat: 3.433	KS-Stat: 0.179	T-stat: 2.797	KS-Stat: 0.108
	P: 0	P: 0.002	P: 0.003	P: 0.154
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
72	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 2.118	KS-Stat: 0.132	T-stat: 4.148	KS-Stat: 0.171
	P: 0.017	P: 0.054	P: 0	P: 0.005
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
84	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.021	KS-Stat: 0.171	T-stat: 3.33	KS-Stat: 0.184
	P: 0.001	P: 0.01	P: 0	P: 0.004
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
96	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.79	KS-Stat: 0.185	T-stat: 3.715	KS-Stat: 0.221
	P: 0	P: 0.006	P: 0	P: 0.001
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
108	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.922	KS-Stat: 0.191	T-stat: 3.538	KS-Stat: 0.265
	P: 0	P: 0.008	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
120	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.937	KS-Stat: 0.242	T-stat: 3.544	KS-Stat: 0.222
	P: 0	P: 0.001	P: 0	P: 0.002

T-test and KS-test results comparing official FTE error distributions obtained via three bins of TVCN GPCE radii. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

Table 8. Statistical Results of TVCN GPCE Radii vs. Official FTE (4 Bins)

Fcst Hour	GPCE Radius vs. Official Error (4 bins)					
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
12	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.541	KS-Stat: 0.162	T-stat: 2.735	KS-Stat: 0.152	T-stat: 4.069	KS-Stat: 0.209
	P: 0	P: 0.003	P: 0.003	P: 0.01	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
24	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 5.574	KS-Stat: 0.252	T-stat: 2.303	KS-Stat: 0.13	T-stat: 3.986	KS-Stat: 0.226
	P: 0	P: 0	P: 0.011	P: 0.047	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
36	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.468	KS-Stat: 0.198	T-stat: 2.97	KS-Stat: 0.188	T-stat: 3.448	KS-Stat: 0.164
	P: 0	P: 0.001	P: 0.002	P: 0.002	P: 0	P: 0.008
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
48	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 3.669	KS-Stat: 0.221	T-stat: 0.713	KS-Stat: 0.073	T-stat: 3.581	KS-Stat: 0.19
	P: 0	P: 0	P: 0.238	P: 0.689	P: 0	P: 0.002
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
60	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 2.175	KS-Stat: 0.157	T-stat: 1.37	KS-Stat: 0.085	T-stat: 2.976	KS-Stat: 0.148
	P: 0.015	P: 0.031	P: 0.086	P: 0.567	P: 0.002	P: 0.05
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
72	T-test: 0	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.508	KS-Stat: 0.153	T-stat: 1.33	KS-Stat: 0.132	T-stat: 4.081	KS-Stat: 0.202
	P: 0.066	P: 0.05	P: 0.092	P: 0.133	P: 0	P: 0.003
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
84	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0
	T-stat: 2.475	KS-Stat: 0.213	T-stat: 2.43	KS-Stat: 0.2	T-stat: 1.477	KS-Stat: 0.153
	P: 0.007	P: 0.004	P: 0.008	P: 0.009	P: 0.07	P: 0.084
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
96	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0
	T-stat: 2.291	KS-Stat: 0.226	T-stat: 2.443	KS-Stat: 0.213	T-stat: 1.781	KS-Stat: 0.146
	P: 0.011	P: 0.003	P: 0.008	P: 0.006	P: 0.038	P: 0.131
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
108	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 1.636	KS-Stat: 0.175	T-stat: 2.436	KS-Stat: 0.191	T-stat: 2.754	KS-Stat: 0.251
	P: 0.052	P: 0.059	P: 0.008	P: 0.033	P: 0.003	P: 0.002
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
120	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 3.192	KS-Stat: 0.301	T-stat: 0.791	KS-Stat: 0.156	T-stat: 3.898	KS-Stat: 0.277
	P: 0.001	P: 0	P: 0.215	P: 0.161	P: 0	P: 0.001

T-test and KS-test results comparing official FTE error distributions obtained via four bins of TVCN GPCE radii. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

Table 9. Statistical Results of TVCN GPCE Radii vs. Official FTE (5 Bins)

Fcst Hour	GPCE Radius vs. Official Error (5 bins)							
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
12	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0
	T-stat: 3.545	KS-Stat: 0.195	T-stat: 1.819	KS-Stat: 0.074	T-stat: 2.35	KS-Stat: 0.211	T-stat: 2.197	KS-Stat: 0.126
	P: 0	P: 0.001	P: 0.035	P: 0.694	P: 0.01	P: 0	P: 0.014	P: 0.139
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
24	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 4.8	KS-Stat: 0.234	T-stat: 1.479	KS-Stat: 0.131	T-stat: 2.235	KS-Stat: 0.154	T-stat: 2.779	KS-Stat: 0.205
	P: 0	P: 0	P: 0.07	P: 0.089	P: 0.013	P: 0.033	P: 0.003	P: 0.002
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
36	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 4.206	KS-Stat: 0.211	T-stat: 1.506	KS-Stat: 0.165	T-stat: 1.748	KS-Stat: 0.095	T-stat: 2.811	KS-Stat: 0.174
	P: 0	P: 0.001	P: 0.067	P: 0.02	P: 0.041	P: 0.496	P: 0.003	P: 0.018
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
48	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0
	T-stat: 3.011	KS-Stat: 0.231	T-stat: -1.04	KS-Stat: 0.135	T-stat: 3.744	KS-Stat: 0.233	T-stat: 1.903	KS-Stat: 0.103
	P: 0.001	P: 0	P: 0.851	P: 0.128	P: 0	P: 0.001	P: 0.029	P: 0.4
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
60	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 0.29	KS-Stat: 0.142	T-stat: 1.153	KS-Stat: 0.093	T-stat: 1.403	KS-Stat: 0.11	T-stat: 2.35	KS-Stat: 0.171
	P: 0.187	P: 0.128	P: 0.125	P: 0.575	P: 0.081	P: 0.391	P: 0.01	P: 0.042
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
72	T-test: 1	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.656	KS-Stat: 0.151	T-stat: 1.373	KS-Stat: 0.119	T-stat: 0.568	KS-Stat: 0.087	T-stat: 3.361	KS-Stat: 0.206
	P: 0.05	P: 0.11	P: 0.086	P: 0.336	P: 0.285	P: 0.73	P: 0	P: 0.012
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
84	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 0
	T-stat: 1.608	KS-Stat: 0.144	T-stat: 0.768	KS-Stat: 0.099	T-stat: 1.885	KS-Stat: 0.174	T-stat: 1.98	KS-Stat: 0.175
	P: 0.055	P: 0.201	P: 0.222	P: 0.648	P: 0.03	P: 0.07	P: 0.025	P: 0.068
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
96	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 0.842	KS-Stat: 0.186	T-stat: 1.118	KS-Stat: 0.17	T-stat: 1.965	KS-Stat: 0.156	T-stat: 2.585	KS-Stat: 0.197
	P: 0.2	P: 0.058	P: 0.133	P: 0.107	P: 0.026	P: 0.165	P: 0.005	P: 0.038
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
108	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: -0.73	KS-Stat: 0.114	T-stat: 3.68	KS-Stat: 0.247	T-stat: 0.86	KS-Stat: 0.119	T-stat: 2.262	KS-Stat: 0.269
	P: 0.765	P: 0.595	P: 0	P: 0.007	P: 0.195	P: 0.529	P: 0.012	P: 0.003
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
120	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 0.605	KS-Stat: 0.122	T-stat: 1.988	KS-Stat: 0.157	T-stat: 0.9	KS-Stat: 0.13	T-stat: 3.648	KS-Stat: 0.284
	P: 0.273	P: 0.548	P: 0.024	P: 0.252	P: 0.185	P: 0.479	P: 0	P: 0.002

T-test and KS-test results comparing official FTE error distributions obtained via five bins of TVCN GPCE radii. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

D. ADDITIONAL FINDINGS

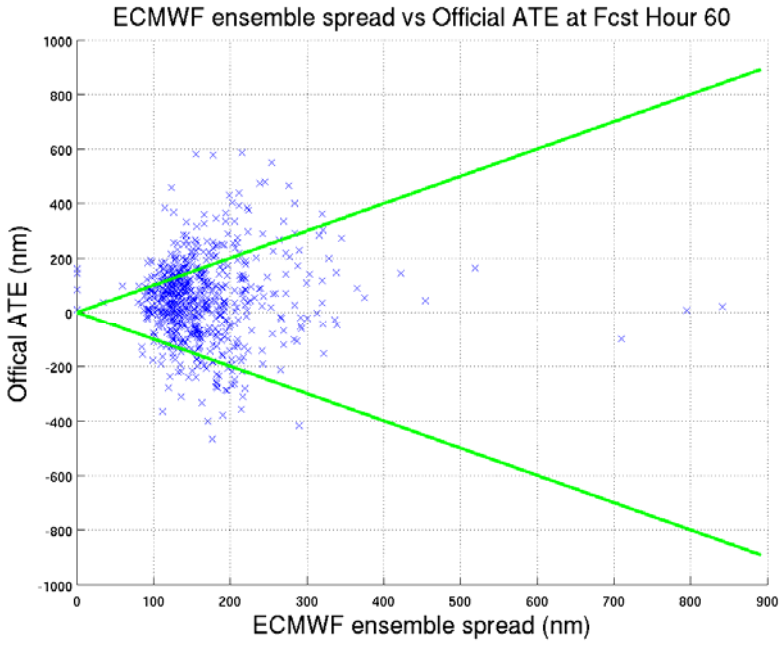
The primary emphasis in this study was to examine the relationship between forecast uncertainty and official FTE. However, a few other possible relationships require analysis as well. The official ATE and XTE are also contained in this data set and may relate to forecast uncertainty differently than the FTE. In addition, the uncertainty measurements may give more definitive information about the errors of the parent model. As such, the ECMWF ensemble spread was compared to official ATE, official XTE, and ECMWF EMN error while TVCN GPCE radii was compared to official ATE, official XTE, and TVCN error.

1. Using ATE and XTE

A similar statistical analysis was performed on the ATE and XTE as was done on FTE. Figures 12 and 13 are scatterplots of ECMWF ensemble spread vs. official ATE and vs. official XTE, respectively. Note that unlike Figures 4 and 5, which use official FTE, not only does the data fail to follow the one-to-one lines, but it does not follow any particular pattern. Visual inspection indicates that there is not a clear relationship between ECMWF ensemble spread and either official ATE or XTE.

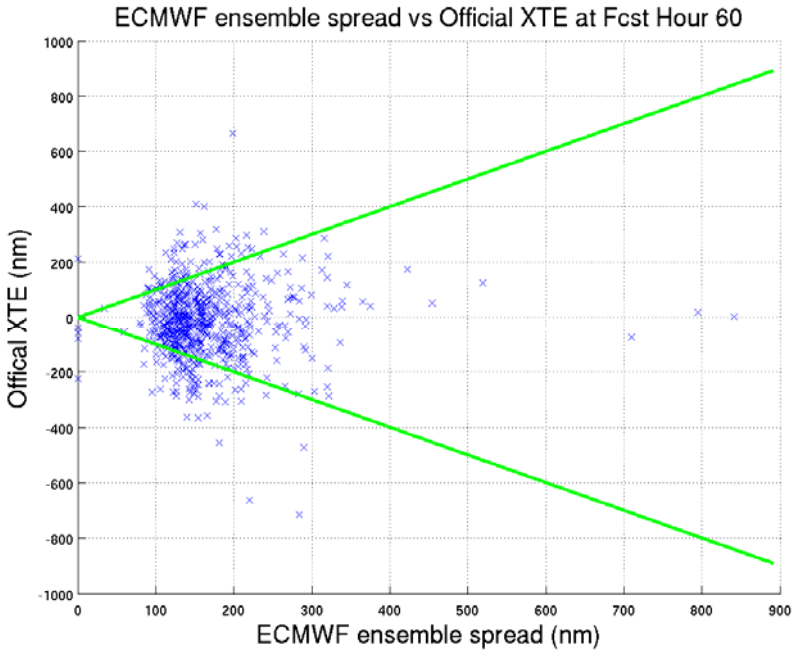
Furthermore, Figures 14 and 15 are scatterplots of TVCN GPCE radii vs. official ATE and vs. official XTE, respectively. Again, the data in both Figures 14 and 15 fails to display any sort of discernable pattern from which a relationship can be established.

Figure 12. ECMWF Ensemble Spread vs. Official ATE at 60 Hours



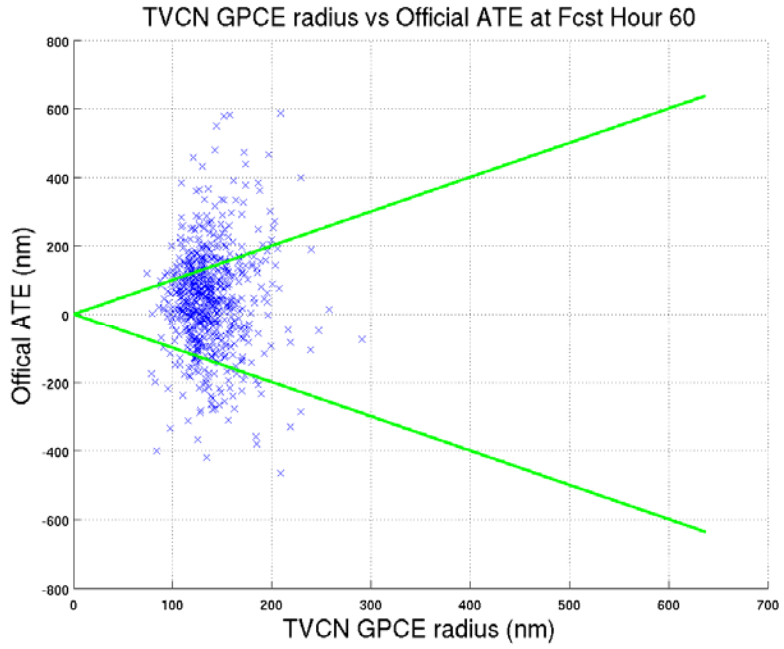
A scatterplot of ECMWF ensemble spread vs. official ATE with positively and negatively sloped one-to-one lines (solid greens).

Figure 13. ECMWF Ensemble Spread vs. Official XTE at 60 Hours



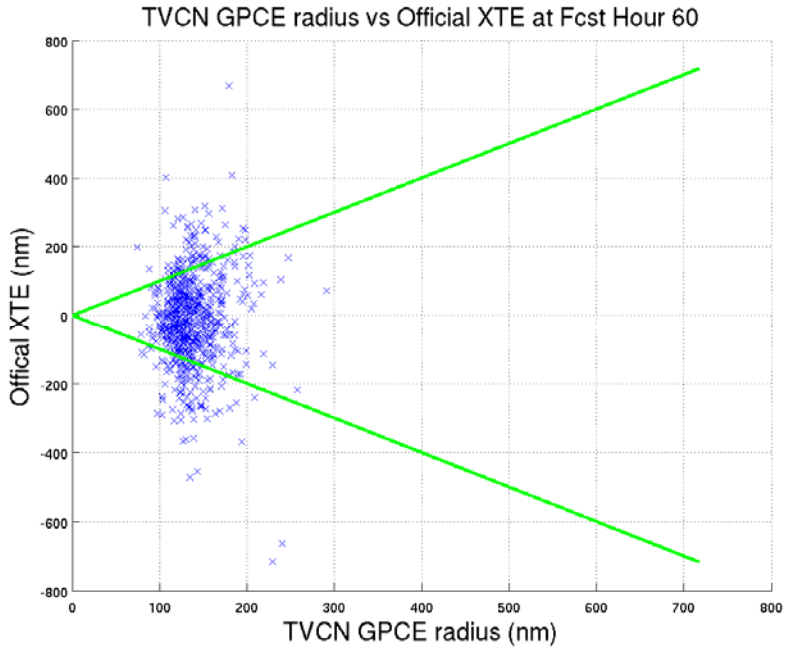
A scatterplot of ECMWF ensemble spread vs. official XTE with positively and negatively sloped one-to-one lines (solid greens).

Figure 14. TVCN GPCE Radii vs. Official ATE at 60 Hours



A scatterplot of TVCN GPCE radii vs. official ATE with positively and negatively sloped one-to-one lines (solid greens).

Figure 15. TVCN GPCE Radii vs. Official XTE at 60 Hours



A scatterplot of TVCN GPCE radii vs. official XTE with positively and negatively sloped one-to-one lines (solid greens).

While visual examination of the scatterplots in Figures 12–15, which plot official ATE and XTE with each uncertainty measurement, does not appear to support any relationships, statistical testing is still required to verify this hypothesis. The t-test and KS-test will be used again to accomplish a more detailed analysis. Similarly to the testing of official FTE, the data will be divided into three bins to begin. However, for this round of tests, the two-tailed t-test will be used since we can no longer assume that each successive bin will have a larger mean error than the previous bin. Due to the fact the ATE and XTE can be positive (representing a forecast too fast or too far right) or negative (representing a forecast too slow or too far left), the mean may not shift far from zero.

a. ECMWF Ensemble Spread vs. Official ATE

This comparison failed to reject the null hypothesis for 29 of 40 tests. All but one test came to this conclusion in the Bin 1 vs. Bin 2 comparison, thus not only did the mean fail to shift, but the distribution shape failed to change as well. These results clear signify that no benefit will be attained by creating even three bins. However, 7 of 10 KS-tests rejected the null hypothesis when comparing Bin 2 to Bin 3. This suggests that the data can be split into two bins which can provide some distinction between the distribution shapes.

b. ECMWF Ensemble Spread vs. Official XTE

This comparison failed to reject the null hypothesis for 30 of 40 tests. All but two tests came to this conclusion in the Bin 1 vs. Bin 2 comparison. As with the previous comparison, this shows that neither the mean or distribution shape changed enough to become statistically different. The Bin 2 vs. Bin 3 comparison shows similar results. Although both the t- and KS-tests reject the null hypothesis for the last three forecast hours, the first half of the forecast range shows no benefit at all. These results also signify that no benefit will be attained by creating three bins, and it appears unlikely that a well-defined relationship can be established from even two bins.

c. TVCN GPCE Radii vs. Official ATE

This comparison failed to reject the null hypothesis for 34 of 40 tests. Neither bin comparison showed very much distinction using either test. These results clearly signify that no benefit will be attained by creating even three bins.

d. TVCN GPCE Radii vs. Official XTE

This comparison failed to reject the null hypothesis for 20 of 40 tests. While this comparison rejected the highest number of null hypotheses out of these last four comparisons, the Bin 1 vs. Bin 2 comparison still lacked much distinction. The majority of distinction was found in the Bin 2 vs. Bin 3 comparison where 14 of 20 tests rejected the null hypothesis (8 of which came from the KS-test). These results indicate that for almost all forecast hours, the error distributions of Bin 2 vs. Bin 3 had different distribution shapes (and different means for most forecast hours). These results give cause for investigating the potential benefit attained by creating two bins, but adding a third bin is clearly not beneficial.

2. ECMWF Ensemble Spread vs. ECMWF EMN Error

In Table 10 are the results from the statistical testing when dividing ECMWF ensemble spread into three bins and comparing each bin's sample mean and distribution of ECMWF EMN errors. Only two tests failed to reject the null hypothesis. This was comparable to the comparisons using official FTE. These results indicate that three ranges ECMWF ensemble spread can successfully provide unique error distributions that are associated with the EMN. This suggests that the relationship between uncertainty and model error (EMN) or official error is similar.

Results from the statistical testing when dividing the ECMWF ensemble spread into four bins and comparing each bin's sample mean and distribution of ECMWF EMN errors are shown in Table 11. Only 8 of 60 tests failed to reject the null hypothesis with only two pairs of tests doing so for the same bin comparison at the same time. These results are the strongest of all methods utilizing four bins. In comparison, ECMWF ensemble spread vs. official FTE failed to reject the null hypothesis for 14 of 60 tests.

This suggests that a better defined relationship may exist between the forecast model uncertainty and error rather than a the model uncertainty and official error.

In Table 12, the results from the statistical testing when dividing the ECMWF ensemble spread into five bins and comparing each bin's sample mean and distribution of ECMWF EMN errors are shown. As indicated in Table 12, 23 of 80 tests failed to reject the null hypothesis; however, three of those tests were within 1% of doing so. These tests include 8 pairs of tests that failed to reject the null hypothesis for the same bin comparison at the same time (only six pairs if tests within 1% are considered). While these results are not concrete, they suggest that there is a good possibility of attaining benefit from using five bins—especially through forecast hour 84. These results also indicate that ECWMF ensemble spread predicts EMN error better than official error at this resolution. This suggests that the ECMWF ensemble spread is more indicative of its own model's performance rather than official forecast performance that utilizes more information than just the ECMWF model.

Table 10. Statistical Results of ECMWF Ensemble Spread vs. EMN Error (3 Bins)

Fcst Hour	ECMWF Spread vs. ECMWF EMN Error (3 bins)			
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
12	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 4.98	KS-Stat: 0.204	T-stat: 5.291	KS-Stat: 0.188
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
24	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.043	KS-Stat: 0.064	T-stat: 8.035	KS-Stat: 0.828
	P: 0.001	P: 0.001	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
36	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 4.742	KS-Stat: 0.216	T-stat: 6.378	KS-Stat: 0.204
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
48	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 4.527	KS-Stat: 0.233	T-stat: 6.892	KS-Stat: 0.284
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
60	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.349	KS-Stat: 0.198	T-stat: 6.04	KS-Stat: 0.24
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
72	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.736	KS-Stat: 0.227	T-stat: 5.798	KS-Stat: 0.256
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
84	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.818	KS-Stat: 0.217	T-stat: 6.281	KS-Stat: 0.299
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
96	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 1.862	KS-Stat: 0.151	T-stat: 6.454	KS-Stat: 0.329
	P: 0.032	P: 0.032	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
108	T-test: 0	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 0.745	KS-Stat: 0.172	T-stat: 7.458	KS-Stat: 0.364
	P: 0.228	P: 0.015	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
120	T-test: 0	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 1.282	KS-Stat: 0.177	T-stat: 6.011	KS-Stat: 0.392
	P: 0.101	P: 0.019	P: 0	P: 0

T-test and KS-test results comparing ECMWF EMN error distributions obtained via three bins of ECMWF ensemble spread. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

Table 11. Statistical Results of ECMWF Ensemble Spread vs. EMN Error (4 Bins)

Fcst Hour	ECMWF Spread vs. ECMWF EMN Error (4 bins)					
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
12	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.918	KS-Stat: 0.171	T-stat: 1.928	KS-Stat: 0.17	T-stat: 4.629	KS-Stat: 0.183
	P: 0	P: 0.002	P: 0.027	P: 0.002	P: 0	P: 0.001
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
24	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 0.673	KS-Stat: 0.112	T-stat: 4.134	KS-Stat: 0.217	T-stat: 6.027	KS-Stat: 0.248
	P: 0.251	P: 0.11	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
36	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.745	KS-Stat: 0.161	T-stat: 3.35	KS-Stat: 0.164	T-stat: 5.154	KS-Stat: 0.198
	P: 0.003	P: 0.008	P: 0	P: 0.007	P: 0	P: 0.001
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
48	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.988	KS-Stat: 0.165	T-stat: 3.252	KS-Stat: 0.206	T-stat: 4.349	KS-Stat: 0.213
	P: 0.001	P: 0.009	P: 0.001	P: 0.001	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
60	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.138	KS-Stat: 0.189	T-stat: 3.613	KS-Stat: 0.197	T-stat: 3.612	KS-Stat: 0.217
	P: 0.017	P: 0.003	P: 0	P: 0.002	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
72	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 3.552	KS-Stat: 0.248	T-stat: 2.509	KS-Stat: 0.134	T-stat: 4.254	KS-Stat: 0.239
	P: 0	P: 0	P: 0.006	P: 0.104	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
84	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 3.256	KS-Stat: 0.271	T-stat: 2.245	KS-Stat: 0.126	T-stat: 4.464	KS-Stat: 0.283
	P: 0.001	P: 0	P: 0.013	P: 0.185	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
96	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.746	KS-Stat: 0.189	T-stat: 1.664	KS-Stat: 0.121	T-stat: 4.981	KS-Stat: 0.341
	P: 0.041	P: 0.015	P: 0.049	P: 0.269	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
108	T-test: 0	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 0.363	KS-Stat: 0.196	T-stat: 2.158	KS-Stat: 0.201	T-stat: 5.377	KS-Stat: 0.292
	P: 0.359	P: 0.016	P: 0.016	P: 0.013	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
120	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: -1.21	KS-Stat: 0.096	T-stat: 2.916	KS-Stat: 0.222	T-stat: 5.625	KS-Stat: 0.33
	P: 0.886	P: 0.678	P: 0.002	P: 0.008	P: 0	P: 0

T-test and KS-test results comparing ECMWF EMN error distributions obtained via four bins of ECMWF ensemble spread. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

Table 12. Statistical Results of ECMWF Ensemble Spread vs. EMN Error (5 Bins)

Fcst Hour	ECMWF Spread vs. ECMWF EMN Error (5 bins)							
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
12	T-test: 1	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 1.943	KS-Stat: 0.108	T-stat: 1.644	KS-Stat: 0.119	T-stat: 1.964	KS-Stat: 0.169	T-stat: 4.865	KS-Stat: 0.237
	P: 0.026	P: 0.208	P: 0.051	P: 0.133	P: 0.025	P: 0.009	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
24	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 0.85	KS-Stat: 0.135	T-stat: 1.755	KS-Stat: 0.142	T-stat: 3.679	KS-Stat: 0.196	T-stat: 4.759	KS-Stat: 0.224
	P: 0.198	P: 0.069	P: 0.04	P: 0.05	P: 0	P: 0.002	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
36	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.801	KS-Stat: 0.121	T-stat: 2.506	KS-Stat: 0.19	T-stat: 2.427	KS-Stat: 0.144	T-stat: 4.138	KS-Stat: 0.225
	P: 0.036	P: 0.167	P: 0.006	P: 0.004	P: 0.008	P: 0.059	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
48	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 1.986	KS-Stat: 0.187	T-stat: 2.224	KS-Stat: 0.155	T-stat: 2.885	KS-Stat: 0.173	T-stat: 3.743	KS-Stat: 0.227
	P: 0.024	P: 0.008	P: 0.013	P: 0.045	P: 0.002	P: 0.018	P: 0	P: 0.001
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
60	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.132	KS-Stat: 0.145	T-stat: 2.032	KS-Stat: 0.191	T-stat: 1.935	KS-Stat: 0.164	T-stat: 3.378	KS-Stat: 0.213
	P: 0.017	P: 0.096	P: 0.022	P: 0.01	P: 0.027	P: 0.041	P: 0	P: 0.003
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
72	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 2.202	KS-Stat: 0.214	T-stat: 1.923	KS-Stat: 0.142	T-stat: 1.919	KS-Stat: 0.16	T-stat: 3.273	KS-Stat: 0.203
	P: 0.014	P: 0.004	P: 0.028	P: 0.137	P: 0.028	P: 0.069	P: 0.001	P: 0.009
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
84	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.141	KS-Stat: 0.237	T-stat: 1.566	KS-Stat: 0.106	T-stat: 2.502	KS-Stat: 0.186	T-stat: 3.86	KS-Stat: 0.253
	P: 0.017	P: 0.002	P: 0.059	P: 0.505	P: 0.007	P: 0.031	P: 0	P: 0.001
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
96	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 0.653	KS-Stat: 0.146	T-stat: 0.826	KS-Stat: 0.075	T-stat: 3.355	KS-Stat: 0.203	T-stat: 2.56	KS-Stat: 0.213
	P: 0.257	P: 0.195	P: 0.205	P: 0.922	P: 0	P: 0.022	P: 0.006	P: 0.014
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
108	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: -0.53	KS-Stat: 0.115	T-stat: 1.097	KS-Stat: 0.142	T-stat: 2.269	KS-Stat: 0.201	T-stat: 5.237	KS-Stat: 0.323
	P: 0.701	P: 0.529	P: 0.137	P: 0.267	P: 0.012	P: 0.035	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
120	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: -2.22	KS-Stat: 0.14	T-stat: 2.152	KS-Stat: 0.221	T-stat: 1.705	KS-Stat: 0.186	T-stat: 5.132	KS-Stat: 0.372
	P: 0.986	P: 0.347	P: 0.016	P: 0.025	P: 0.045	P: 0.09	P: 0	P: 0

T-test and KS-test results comparing ECMWF EMN error distributions obtained via five bins of ECMWF ensemble spread. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

3. TVCN GPCE Radii vs. TVCN Error

In Table 13 are shown the results from the statistical testing when dividing TVCN GPCE radii into three bins and comparing each bin's sample mean and distribution of TVCN errors. Only two tests failed to reject the null hypothesis, and one of those was within 1%. These results, similarly to all other comparisons except those involving ATE and XTE, indicate that three ranges of TVCN GPCE radii can successfully predict unique error distributions for the TVCN.

Shown in Table 14 are the results from the statistical testing when dividing TVCN GPCE radii into four bins and comparing each bin's sample mean and distribution of TVCN errors. Only 9 of 60 tests failed to reject the null hypothesis (one within 1%). Of those, only two pairs of tests failed to reject the null hypothesis for the same bin comparison at the same forecast hour. These results indicate that establishing four bins of GPCE radii can add benefit to the TVCN consensus forecast. The strength of this relationship is comparable to that between ECMWF ensemble spread vs. EMN error and slightly better than that between GPCE radii vs. official FTE.

Results from the statistical testing when dividing TVCN GPCE radii into five bins and comparing each bin's sample mean and distribution of TVCN errors are shown in Table 15. When utilizing five bins, 34 of 80 tests fail to reject the null hypothesis with 13 pairs of tests doing so for the same bin comparison and forecast hour. Furthermore, half of the forecast hours only maintain distinction between two of the bin comparisons. These results indicate that four bins was the maximum supported by this data pool. While these results are similar to those found between GPCE vs. official FTE, this relationship is not as strong as that found between ECMWF ensemble spread vs. EMN error. This suggests ECMWF ensemble spread may be a better predictor of error than TVCN GPCE radii.

Table 13. Statistical Results of TVCN GPCE Radii vs. TVCN Error (3 Bins)

Fcst Hour	GPCE Radius vs. TVCN Error (3 bins)			
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
12	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 5.137	KS-Stat: 0.205	T-stat: 6.353	KS-Stat: 0.247
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
24	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 5.588	KS-Stat: 0.222	T-stat: 6.409	KS-Stat: 0.265
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
36	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 5.525	KS-Stat: 0.229	T-stat: 5.807	KS-Stat: 0.233
	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
48	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.967	KS-Stat: 0.13	T-stat: 6.264	KS-Stat: 0.271
	P: 0.002	P: 0.023	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
60	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.339	KS-Stat: 0.16	T-stat: 4.828	KS-Stat: 0.189
	P: 0	P: 0.005	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
72	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.57	KS-Stat: 0.094	T-stat: 5.886	KS-Stat: 0.263
	P: 0.059	P: 0.228	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
84	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.021	KS-Stat: 0.14	T-stat: 4.012	KS-Stat: 0.243
	P: 0.001	P: 0.043	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
96	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.657	KS-Stat: 0.186	T-stat: 4.245	KS-Stat: 0.263
	P: 0	P: 0.004	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
108	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 4.289	KS-Stat: 0.214	T-stat: 4.416	KS-Stat: 0.301
	P: 0	P: 0.001	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3	
120	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 4.819	KS-Stat: 0.252	T-stat: 4.553	KS-Stat: 0.254
	P: 0	P: 0	P: 0	P: 0

T-test and KS-test results comparing TVCN error distributions obtained via three bins of TVCN GPCE radii. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

Table 14. Statistical Results of TVCN GPCE Radii vs. TVCN Error (4 Bins)

Fcst Hour	GPCE Radius vs. TVCN Error (4 bins)					
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
12	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.612	KS-Stat: 0.172	T-stat: 3.611	KS-Stat: 0.178	T-stat: 4.629	KS-Stat: 0.217
	P: 0	P: 0.001	P: 0	P: 0.002	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
24	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 5.193	KS-Stat: 0.248	T-stat: 2.428	KS-Stat: 0.141	T-stat: 4.813	KS-Stat: 0.261
	P: 0	P: 0	P: 0.008	P: 0.028	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
36	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.52	KS-Stat: 0.141	T-stat: 3.905	KS-Stat: 0.209	T-stat: 4.689	KS-Stat: 0.223
	P: 0	P: 0.026	P: 0	P: 0	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
48	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 3.398	KS-Stat: 0.168	T-stat: 1.013	KS-Stat: 0.096	T-stat: 5.776	KS-Stat: 0.26
	P: 0	P: 0.007	P: 0.156	P: 0.33	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
60	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 2.143	KS-Stat: 0.157	T-stat: 2.023	KS-Stat: 0.111	T-stat: 4.153	KS-Stat: 0.204
	P: 0	P: 0.024	P: 0.022	P: 0.237	P: 0	P: 0.001
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
72	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.608	KS-Stat: 0.13	T-stat: 1.394	KS-Stat: 0.129	T-stat: 5.131	KS-Stat: 0.269
	P: 0.054	P: 0.118	P: 0.082	P: 0.139	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
84	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.053	KS-Stat: 0.169	T-stat: 2.733	KS-Stat: 0.182	T-stat: 2.076	KS-Stat: 0.206
	P: 0.02	P: 0.029	P: 0.003	P: 0.015	P: 0.019	P: 0.004
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
96	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 1
	T-stat: 2.376	KS-Stat: 0.196	T-stat: 2.895	KS-Stat: 0.217	T-stat: 1.532	KS-Stat: 0.215
	P: 0.009	P: 0.01	P: 0.002	P: 0.003	P: 0.063	P: 0.004
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
108	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.874	KS-Stat: 0.27	T-stat: 2.458	KS-Stat: 0.203	T-stat: 2.904	KS-Stat: 0.222
	P: 0.002	P: 0	P: 0.007	P: 0.014	P: 0.002	P: 0.005
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4	
120	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 3.563	KS-Stat: 0.201	T-stat: 2.021	KS-Stat: 0.13	T-stat: 3.547	KS-Stat: 0.274
	P: 0	P: 0.024	P: 0.022	P: 0.315	P: 0	P: 0.001

T-test and KS-test results comparing TVCN error distributions obtained via four bins of TVCN GPCE radii. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

Table 15. Statistical Results of TVCN GPCE radii vs. TVCN Error (5 Bins)

Fcst Hour	GPCE Radius vs. TVCN Error (5 bins)							
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
12	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 3.721	KS-Stat: 0.216	T-stat: 2.019	KS-Stat: 0.086	T-stat: 2.87	KS-Stat: 0.208	T-stat: 2.987	KS-Stat: 0.1543
	P: 0	P: 0	P: 0.022	P: 0.509	P: 0.002	P: 0.001	P: 0.002	P: 0.03
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
24	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 5.116	KS-Stat: 0.25	T-stat: 1.05	KS-Stat: 0.105	T-stat: 2.832	KS-Stat: 0.164	T-stat: 3.53	KS-Stat: 0.197
	P: 0	P: 0	P: 0.147	P: 0.272	P: 0.002	P: 0.017	P: 0	P: 0.002
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
36	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 3.68	KS-Stat: 0.178	T-stat: 1.734	KS-Stat: 0.145	T-stat: 1.888	KS-Stat: 0.119	T-stat: 4.436	KS-Stat: 0.248
	P: 0	P: 0.008	P: 0.042	P: 0.064	P: 0.03	P: 0.204	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
48	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 2.922	KS-Stat: 0.152	T-stat: -1.03	KS-Stat: 0.094	T-stat: 4.092	KS-Stat: 0.241	T-stat: 4.087	KS-Stat: 0.209
	P: 0.002	P: 0.046	P: 0.849	P: 0.477	P: 0	P: 0	P: 0	P: 0.003
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
60	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 1	KS-Test: 1
	T-stat: 0.76	KS-Stat: 0.114	T-stat: 1.239	KS-Stat: 0.093	T-stat: 2.082	KS-Stat: 0.179	T-stat: 3.657	KS-Stat: 0.204
	P: 0.224	P: 0.294	P: 0.108	P: 0.577	P: 0.019	P: 0.023	P: 0	P: 0.005
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
72	T-test: 1	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.694	KS-Stat: 0.125	T-stat: 0.726	KS-Stat: 0.075	T-stat: 1.511	KS-Stat: 0.138	T-stat: 4.039	KS-Stat: 0.261
	P: 0.046	P: 0.253	P: 0.234	P: 0.856	P: 0.066	P: 0.171	P: 0	P: 0
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
84	T-test: 1	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.715	KS-Stat: 0.156	T-stat: 1.173	KS-Stat: 0.114	T-stat: 1.54	KS-Stat: 0.135	T-stat: 1.88	KS-Stat: 0.244
	P: 0.044	P: 0.105	P: 0.121	P: 0.43	P: 0.062	P: 0.243	P: 0.031	P: 0.002
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
96	T-test: 0	KS-Test: 0	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.363	KS-Stat: 0.171	T-stat: 1.339	KS-Stat: 0.146	T-stat: 2.024	KS-Stat: 0.172	T-stat: 1.899	KS-Stat: 0.231
	P: 0.087	P: 0.082	P: 0.091	P: 0.198	P: 0.022	P: 0.084	P: 0.03	P: 0.006
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
108	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 0.112	KS-Stat: 0.117	T-stat: 3.966	KS-Stat: 0.268	T-stat: 1.324	KS-Stat: 0.115	T-stat: 1.978	KS-Stat: 0.223
	P: 3.456	P: 0.515	P: 0	P: 0.002	P: 0.094	P: 0.543	P: 0.025	P: 0.017
	Bin 1 vs. Bin 2		Bin 2 vs. Bin 3		Bin 3 vs. Bin 4		Bin 4 vs. Bin 5	
120	T-test: 0	KS-Test: 0	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 0	T-test: 1	KS-Test: 1
	T-stat: 1.613	KS-Stat: 0.148	T-stat: 1.891	KS-Stat: 0.161	T-stat: 2.384	KS-Stat: 0.182	T-stat: 2.59	KS-Stat: 0.294
	P: 0.054	P: 0.291	P: 0.03	P: 0.198	P: 0.009	P: 0.108	P: 0.005	P: 0.001

T-test and KS-test results comparing TVCN error distributions obtained via five bins of TVCN GPCE radii. Green cells indicate that the null hypothesis is rejected. Red cells indicate a failure to reject the null hypothesis. T-stat = T-statistic, KS-Stat = KS-statistic, and P = P-value.

V. CONCLUSION AND RECOMMENDATIONS

A. CONCLUSIONS

Since the inherent error present in all forecasts will never be eliminated, forecasters must do their best to characterize and quantify the associated uncertainty with weather forecasts—especially TC forecasts which directly influence decisions that affect millions of people and billions of dollars. This thesis aimed to improve the tools which TC forecasters at NHC use operationally. This was accomplished by creating the maximum number of bins (ranges of uncertainty measurements) that would contain unique error distributions as measured by mean and shape. These bins and associated error distributions could then be utilized by the MC method where 1,000 errors are pulled and applied to NHC's official forecast. The result is a tailored TC track forecast with improved estimates of uncertainty for each forecast disseminated. In the end, these improvements directly advance the NHC WSP products which are used by decision makers to mitigate TC impacts.

This thesis found that the maximum number of bins that will still maintain unique error distributions are as follows:

- For ECMWF ensemble spread (using official FTE): 4 bins
- For ECMWF ensemble spread (using ECMWF EMN error): 5 bins
- For TVCN GPCE radii (using official FTE): 4 bins
- For TVCN GPCE radii (using TVCN error): 4 bins

This thesis also found that using official ATE and XTE to populate the same bins of uncertainty did not produce statistically different distributions at even the three bin level. This suggests that official ATE and XTE are not very well related with either measurement of uncertainty analyzed in this work.

Although the desire to develop a continuous uncertainty-error relationship was not fully realized, the results of this thesis are still promising. NHC currently uses this approach but with only three bins. This research paves the way for testing the possibility of expanding to four bins of uncertainty for operational use.

Another promising finding from this research is that using ECMWF ensemble spread as the measurement of uncertainty coupled with the errors produced from the ECMWF EMN, produced the strongest relationship and showed additional benefit out to five bins. While the original purpose of this thesis was to establish such a relationship with official FTE so that the MC method could be better applied to each new official TC forecast, this alternative relationship can still greatly aid NHC forecasters. By running the MC method on the ECMWF EMN forecast, the already superior ECMWF model output can be further improved to be used as a predominate tool during the creation of official forecasts.

B. RECOMMENDATIONS

While the longer data sample spanning nine years and the use of ECMWF spread show promise that a more continuous uncertainty-error relationship can be derived, there are several things that might improve these results. Future research should investigate:

- Different methods of measuring and/or binning the measurements of uncertainty based on some other storm characteristic (e.g. intensity)
- Omitting outliers in the data, or determining which are most problematic
- Filtering data to exclude errors produced from model runs with a limited number of ensemble members
- Splitting uncertainty measurements into along- and cross-track values and testing them against ATE and XTE.

APPENDIX. RANGES FOR BINS OF UNCERTAINTY

Table 16. ECMWF Ensemble Spread Ranges (3 Bins)

Ranges of ECMWF Ensemble Spread for 3 Bins (nm)			
Fcst Hour	Bin 1	Bin 2	Bin 3
12	≤ 30	31 - 41	> 41
24	≤ 53	54 - 102	> 102
36	≤ 79	80 - 102	> 102
48	≤ 106	107 - 134	> 134
60	≤ 132	133 - 169	> 169
72	≤ 155	156 - 205	> 205
84	≤ 183	184 - 242	> 242
96	≤ 210	211 - 285	> 285
108	≤ 236	237 - 319	> 319
120	≤ 264	265 - 360	> 360

Table 17. ECMWF Ensemble Spread Ranges (4 Bins)

Ranges of ECMWF Ensemble Spread for 4 Bins (nm)				
Fcst Hour	Bin 1	Bin 2	Bin 3	Bin 4
12	≤ 28	29 - 35	36 - 46	> 46
24	≤ 50	51 - 60	61 - 75	> 75
36	≤ 74	75 - 89	90 - 108	> 108
48	≤ 100	101 - 118	119 - 146	> 146
60	≤ 125	126 - 147	148 - 185	> 185
72	≤ 147	148 - 180	181 - 226	> 226
84	≤ 168	169 - 209	210 - 261	> 261
96	≤ 196	197 - 239	240 - 314	> 314
108	≤ 218	219 - 271	272 - 355	> 355
120	≤ 245	246 - 315	316 - 418	> 418

Table 18. ECMWF Ensemble Spread Ranges (5 Bins)

Ranges of ECMWF Ensemble Spread for 5 Bins (nm)					
Fcst Hour	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
12	≤ 27	28 - 32	33 - 38	39 - 49	> 49
24	≤ 47	48 - 56	57 - 66	67 - 79	> 79
36	≤ 71	72 - 83	84 - 97	98 - 114	> 114
48	≤ 97	98 - 111	112 - 127	128 - 154	> 154
60	≤ 121	122 - 137	138 - 159	160 - 196	> 196
72	≤ 142	143 - 166	167 - 195	196 - 237	> 237
84	≤ 162	163 - 192	193 - 226	227 - 276	> 276
96	≤ 185	186 - 221	222 - 267	268 - 331	> 331
108	≤ 207	208 - 251	252 - 298	299 - 390	> 390
120	≤ 235	236 - 284	285 - 339	340 - 441	> 441

Table 19. TVCN GPCE Radii Ranges (3 Bins)

Ranges of TVCN GPCE Radius for 3 Bins (nm)			
Fcst Hour	Bin 1	Bin 2	Bin 3
12	≤ 33	34 - 39	> 39
24	≤ 55	56 - 64	> 64
36	≤ 76	77 - 89	> 89
48	≤ 97	98 - 114	> 114
60	≤ 124	125 - 143	> 143
72	≤ 150	151 - 173	> 173
84	≤ 182	183 - 211	> 211
96	≤ 213	214 - 247	> 247
108	≤ 251	252 - 296	> 296
120	≤ 287	288 - 345	> 345

Table 20. TVCN GPCE Radii Ranges (4 Bins)

Ranges of TVCN GPCE Radius for 4 Bins (nm)				
Fcst Hour	Bin 1	Bin 2	Bin 3	Bin 4
12	≤ 31	32 - 36	37 - 41	> 41
24	≤ 51	52 - 59	60 - 67	> 67
36	≤ 72	73 - 82	83 - 94	> 94
48	≤ 93	94 - 104	105 - 119	> 119
60	≤ 120	121 - 132	133 - 150	> 150
72	≤ 144	145 - 160	161 - 182	> 182
84	≤ 176	177 - 194	195 - 224	> 224
96	≤ 203	204 - 227	228 - 265	> 265
108	≤ 241	242 - 271	272 - 316	> 316
120	≤ 276	277 - 311	312 - 370	> 370

Table 21. TVCN GPCE Radii Ranges (5 Bins)

Ranges of TVCN GPCE Radius for 5 Bins (nm)					
Fcst Hour	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
12	≤ 30	31 - 34	35 - 38	39 - 42	> 42
24	≤ 49	50 - 56	57 - 62	63 - 69	> 69
36	≤ 69	70 - 79	80 - 87	88 - 96	> 96
48	≤ 90	91 - 100	101 - 109	110 - 123	> 123
60	≤ 117	118 - 128	129 - 139	140 - 156	> 156
72	≤ 142	143 - 154	155 - 168	169 - 190	> 190
84	≤ 170	171 - 188	189 - 203	204 - 234	> 234
96	≤ 198	199 - 218	219 - 239	240 - 277	> 277
108	≤ 234	235 - 259	260 - 287	288 - 327	> 327
120	≤ 268	269 - 296	297 - 331	332 - 390	> 390

LIST OF REFERENCES

DeMaria, M., J. Knaff, R. Kanbb, C. Lauer, C. Sampson, and R. DeMaria, 2009: A new method for estimating tropical cyclone wind speed probabilities. *Wea. Forecasting,* **24,** 1573–1591, doi:10.1175/2009WAF2222286.1.

DeMaria, M., and Coauthors, 2013: Improvements to the operational tropical cyclone wind speed probability model, *Wea. Forecasting,* **28,** 586–602, doi:10.1175/WAF-D-12–00116.1.

Goerss, J., 2007: Prediction of consensus tropical cyclone track forecast error. *Mon. Wea. Rev.,***135,** 1985–1993, doi:10.1175/MWR3390.1.

Hauke, M., 2006: Evaluating Atlantic tropical cyclone track error distributions based on forecast confidence. M.S. thesis, Dept. of Meteorology, Naval Postgraduate School, 84 pp.

Neese, J., 2010: Evaluating Atlantic tropical cyclone track error distributions for use in probabilistic forecasts of wind distribution. M.S. thesis, Dept. of Meteorology, Naval Postgraduate School, 87 pp.

NHC, 2014: Tropical cyclone wind speed probabilities products. Accessed on 19 January 2016. [Available online at http://www.nhc.noaa.gov/about/pdf/About_Windspeed_Probabilities.pdf.]

Pearman, D., 2011: Evaluating tropical cyclone forecast track uncertainty using a grand ensemble of ensemble prediction systems. M.S. thesis, Dept. of Meteorology, Naval Postgraduate School, 63 pp.

Santoalla, D., 2015: TIGGE. ECMWF, Accessed on 14 March 2016. [Available online at https://software.ecmwf.int/wiki/display/TIGGE]

Scherrer, S., 2002: Skill Prediction for Medium-Range Weather Forecasts. Diploma thesis, Swiss Federal Institute of Technology Zurich Switzerland, 84 pp.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California